UNIVERSITAT DE BARCELONA

Identification of new candidate genes for germline predisposition to familial colorectal cancer using somatic mutational profiling

Marcos Díaz Gay

Identification of new candidate genes for germline predisposition to familial colorectal cancer using somatic mutational profiling

Manuscript submitted by:
Marcos Díaz Gay

For the degree of:
Doctor in Medicine and Translational Research by the University of Barcelona

Doctoral thesis done in:
Genetic Predisposition to Gastrointestinal Cancer Group
Gastrointestinal and Pancreatic Oncology Team
Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

Directed by:
Dr. Sergi Castellví i Bel
Dr. Francesc Balaguer Prunés

Barcelona, 2019

THESIS DIRECTOR
Dr. Sergi Castellví i Bel

THESIS DIRECTOR
Dr. Francesc Balaguer Prunés

THESIS TUTOR
Dr. Antoni Castells i Garangou

CANDIDATE
Marcos Díaz Gay

IDIBAPS

UNIVERSITAT DE BARCELONA

A meus pais e a Toño

It matters not how strait the gate,
How charged with punishments the scroll,
I am the master of my fate:
I am the captain of my soul.

William Ernest Henley (1849-1903)

Acknowledgements

Todavía recuerdo como si fuese hoy esa entrevista. La entrevista. La que me cambió la vida para siempre. No sé qué visteis en ese Ingeniero de Caminos, Canales y Puertos de 24 años que decía que soñaba con ser científico. Supongo que muchas ganas y mucha motivación, pero no deja de ser arriesgado. Por eso, y por todo lo demás que ha venido en estos 4 años, mi primer agradecimiento es para ti Sergi. Por la enorme confianza que pusiste en mí desde el primer momento. Por darme la oportunidad de cumplir un sueño que supongo que con esta tesis todavía acaba de empezar. Gracias también Francesc por tu ayuda, especialmente en esta fase final de la tesis.

Gracias también a las dos personas que estaban junto al jefe en esa mesa de reuniones de la cuarta planta, gracias a Clara y en especial a Sebas, mis primeros guías y mentores en esta aventura. Gracias Sebas por todos los momentos que hemos compartido durante nuestro viaje por la tesis doctoral, especialmente por todas las risas dentro y fuera del lab. El futuro americano se presenta apasionante. Let's go for it!

Gracias a toda la gente del laboratorio. A mi familia. A los que habéis hecho que esta aventura haya tenido sentido y que haya sido tan bonita y llena de momentos felices.

A los que un día torturé un poquito haciendo de profe de bioinformática y que con el tiempo se convirtieron en grandes amigos de los que he aprendido mucho, Anna, Roser y especialmente mis queridos Paula y Mariano. Espero ir a visitaros pronto al otro lado del mundo.

A los técnicos, que cada día están ahí levantando el laboratorio y haciéndonos todo más fácil al resto. Por una labor que por veces no es demasiado valorada, pero que he apreciado mucho cuando la he necesitado. Elena, Manuel, Esther, Agatha, Nereida, Carolina y en particular mi gran vecina de mesa durante 4 años Jenny.

Al resto de jefes de grupo del laboratorio, Jordi, Txell, Esther y Juanjo, dispuestos a echar una mano y hacer del lab un sitio mejor en el que trabajar.

A la vieja guardia de bioinformáticos, que un día dominabais la -1 junto al infiltrado que siempre habéis tenido en la cuarta, Guillaume, Guerau, Dani, Pau y en especial Maria, artífice también de que buena parte de esta tesis sea real. Y pensar que todo salió de un café mañanero de los nuestros en el Salzburg…

A la nueva generación de doctorandos, Elena, Sara, Javi, Cristina y Yasmin, porque sé que el lab queda en buenas manos y que haréis lo máximo por mantener el buen ambiente que siempre hemos disfrutado.

A la OG, la Old Generation (que por fin tenemos un nombre cool!). Porque sois los máximos culpables de haber creado el ambiente que ha caracterizado nuestro grupo durante todos estos años. Porque cualquier mal día, experimento fallido o programa que no funcionaba nada podían hacer contra unas cervecitas en la china, una buena cenita o simplemente unas palabras de apoyo con unos cafecitos en la sala de reuniones. Es un absoluto placer poder decir que he compartido este viaje del doctorado con un grupo de personas tan increíble como vosotros. Por cada sonrisa que me habéis sacado entre cenas, rafting, karaokes, esquiadas, salidas y demás. Moltíssimes gràcies, Claudia, María, Saray, Lorena, Elena, Irene, Isa, Coral, Laia, Eva, Sebas, Maria i Keyvan.

Gracias también a todas las personas que habéis hecho de Barcelona un lugar mejor para mí durante las distintas etapas que he tenido la oportunidad de vivir en estos años, a los que ya me traía de casa y a los que he tenido la oportunidad de conocer durante el camino, en particular a Laura, Conde, Hache, Diñei, Silvia, Dani, Ari, Manu, María, Fer, Cris, Nave, mis PhD Warriors (Núria, Mire, Martí e Íñigo), la gente de La Huella Aribau y especialmente Candela.

Thanks so much also to my San Diego people. Ludmil, thanks for the opportunity to join your amazing lab at UCSD and making me feel at home during the 4 months that again changed my life forever. I am really grateful to all the people in the lab, Mishu, Burçak, Maria, Erik, Ashrith, Phoebe, Chris, Evelyn, Arielle, Frances, Jason, Nora, George, Adam and Azhar; and also to my San Diego crew outside the lab, David, Alberto, the CrossFit Invictus people, Robert, Ramón, Núria, Josh, Raphael and especially my roommates Dominik, Lukas and Brennan.

E, por suposto, gracias aos meus. Á miña xente de casa. A aqueles que sempre están aí, pese á distancia. Por cada palabra de apoio. Por cada momento compartido, case sempre rodeado de sonrisas e felicidade. Quérovos moitísimo. Gracias a Figueiras, Davis, Ánder, á xente de Santiago e Coruña, e especialmente ao meu querido comando raxo, Hache e Ana, e aos meus pros, que cada día están aí e sei que podo contar sempre con eles, de corazón moitas gracias por ser como sodes, Conde, Dani, Chino, Santi, Pabs, Suso e o meu *primo* de Terrassa Diñei.

Por último, gracias tamén á miña familia. Gracias de corazón especialmente a meus pais e ao meu querido Toño, fonte do meu maior apoio e cariño cada día. Esta tesis non sería posible sen a vosa axuda. Gracias por coidarme tanto e facerme sentir tan afortunado. Por axudarme sempre e ensinarme a que son eu o que dirixe o meu destino... *I am the master of my fate: I am the captain of my soul*.

Summary

Colorectal cancer (CRC) is one of the malignant neoplasms with higher incidence and mortality in Spain, Europe and worldwide. As a complex disease, both environmental and genetic factors influence CRC predisposition. Up to 35% of CRC patients present familial aggregation for the disease, whereas only around 2-8% of cases are linked to a well-known hereditary syndrome associated to pathogenic germline alterations in specific genes, namely *APC*, *MUTYH*, *POLE*, *POLD1* or the DNA mismatch repair genes. During last years, next generation sequencing (NGS) techniques such as whole exome sequencing (WES) have been used to address this gap of missing heritability. Characterization of somatic mutational profiles, performed by the application of NGS to both germline and tumor DNA, has also been recently established as a powerful tool to identify novel genes linked to CRC predisposition. However, although some bioinformatic packages have been developed to address this analysis, it remains inaccessible for a substantial proportion of the scientific community. Accordingly, the main purpose of this doctoral thesis was to identify new genes involved in germline predisposition to familial CRC, by using an integrated germline-tumor WES analysis and somatic mutational profiling, as well as facilitating the application of these genomic analyses to the scientific community.

As a first step, a bioinformatic tool to deal with somatic mutational profiling was developed. Shiny framework was used to build MuSiCa, a user-friendly web application freely accessible and potentially useful for non-specialized researchers. Tumor mutational burden calculation and mutational signature refitting analysis according to the information present in COSMIC database is available, as well as different options for sample classification through clustering and principal component analysis.

Subsequently, an integrated germline-tumor analysis was implemented in a cohort of 18 familial CRC unrelated patients. WES data of both germline and tumor DNA was available, allowing the identification of new potential tumor suppressor genes according to Knudson's two-hit hypothesis. Benefitting from the development of MuSiCa application, somatic mutational profiling was also analyzed, uncovering five hypermutated samples. An enrichment of DNA repair-associated genes was found, as well as some genes previously linked to predisposition syndromes to other cancer types. *BRCA2*, *BLM*, *ERCC2*, *RECQL*, *REV3L* and *RIF1* were found as the most promising candidate genes for germline CRC predisposition. Interestingly, a germline mutation was found in the DNA repair gene *RECQL* in a patient with one of the hypermutated tumors, reinforcing the putative role of this gene in hereditary CRC. These findings could be helpful in clinical practice improving genetic counseling in the affected families.

Contents

Abbreviations

APOBEC	Apolipoprotein B mRNA editing catalytic polypeptide-like
BER	Base excision repair
Cas9	CRISPR-associated protein 9
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CMMRD	Constitutional mismatch repair deficiency
CMS	Consensus molecular subtype
CNA	Copy number alteration
CNV	Copy number variant
COSMIC	Catalogue of somatic mutations in cancer
CRC	Colorectal cancer
CRISPR	Clustered regularly interspaced short palindromic repeats
CTLA-4	Cytotoxic T-lymphocyte-associated antigen 4
DBS	Doublet base substitution
DNA	Deoxyribonucleic acid
EMAST	Elevated microsatellite alterations at selected tetranucleotide repeats
ENCODE	Encyclopedia of DNA Elements
ExAC	Exome Aggregation Consortium
FAP	Familial adenomatous polyposis
GAPPS	Gastric adenocarcinoma and proximal polyposis of the stomach
GATK	Genome analysis toolkit
gnomAD	Genome Aggregation Database
GUI	Graphical user interface
GWAS	Genome wide association studies
Indel	Small insertion or deletion
LOH	Loss of heterozygosity
MAF	Mutation annotation format
MAP	*MUTYH*-associated polyposis
MLPA	Multiplex ligation-dependent probe amplification
MMR	Mismatch repair
MSI	Microsatellite instability
MuSiCa	Mutational signatures in cancer
NATS	*NTHL1*-associated tumor syndrome
NER	Nucleotide excision repair
NGS	Next generation sequencing

NMF	Non-negative matrix factorization
NNLS	Non-negative least squares
PD-1	Programmed cell death protein 1
PPAP	Polymerase proofreading-associated polyposis
ROS	Reactive oxygen species
SBS	Single base substitution
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPS	Serrated polyposis syndrome
SV	Structural variant
TCGA	The cancer genome atlas
TMB	Tumor mutational burden
TSG	Tumor suppressor gene
TSV	Tab-separated values
UCC	Urothelial cell cancer
UPD	Uniparental disomy
UV	Ultraviolet
VCF	Variant call format
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World health organization

INTRODUCTION

1. Colorectal cancer

1.1 Epidemiology

Colorectal cancer (CRC) is one of the most common malignant neoplasms worldwide with a significant associated mortality. According to data from the International Agency for Research on Cancer, more than a million and a half new cases are diagnosed and over 800,000 people die from this pathology each year, accounting for 9% of all cancer-related deaths. Among all cancer sites and considering both genders, CRC ranks third regarding incidence but second with respect to mortality (**Figure 1**) (Bray et al., 2018).

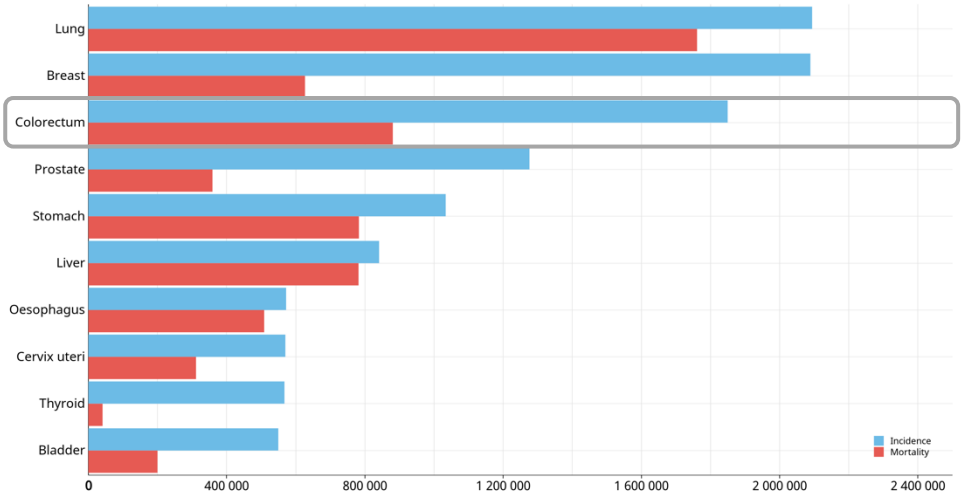


Figure 1. Number of incident cancer cases and deaths worldwide considering both genders and all ages. Top 10 cancer sites ordered by worldwide number of incidence cases according to data from GLOBOCAN, produced by the International Agency for Research on Cancer. Gray box highlights colorectal cancer cases (Ferlay et al., 2019).

One in 37 men and 1 in 55 women will develop the disease and 1 in 88 men and 1 in 139 women will die from it before age 75 years. Regarding geographical distribution, the highest incidence ratios are found in Australia and New Zealand, Europe, Eastern Asia and North America, all above 25 new cases per 100,000 persons per year (**Figure 2**) (Ferlay et al., 2019). Survival rates varied widely, although they are lower in low-income countries (Brenner, Kloor, & Pox, 2014; Allemani et al., 2018). Early detection plays a big role in survival, since 90% of patients diagnosed in the localized stage survive after 5 years from diagnosis, whereas only around 10% of those diagnosed with distant tumor spread (Siegel et al., 2017; Brenner & Chen, 2018).

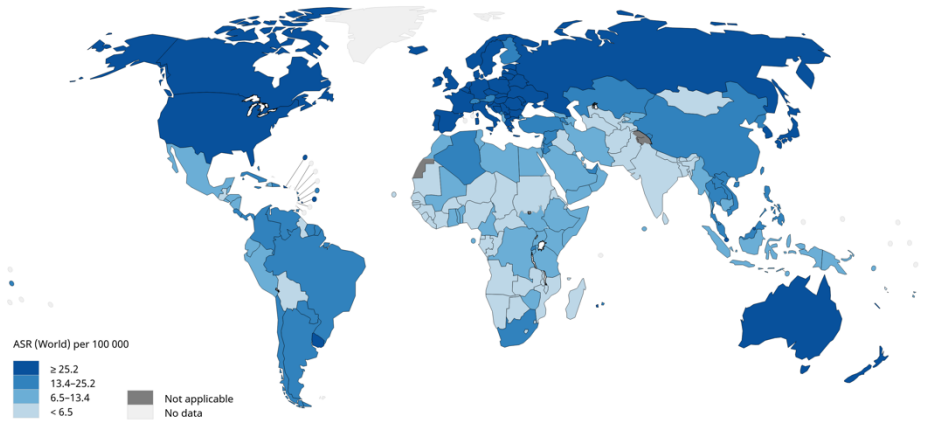


Figure 2. Colorectal cancer incidence rates. Age and population-standardized worldwide incidence rates in 2018 according to GLOBOCAN (Ferlay et al., 2019).

A correlation between cancer incidence and socioeconomical development was previously observed for all cancers by comparing the Human Development Index (measuring life expectancy, education and economic development of a given country) (United Nations Development Programme, 2019) against the ranking of cancer as a premature cause of death (before age 70 years) (Bray et al., 2018). This fact is in agreement with CRC incidence being higher in more developed regions (Ferlay et al., 2019). However, these figures could be drastically influenced by the quality of the medical registries, resulting in a huge number of cancer patients being ignored by their countries health services (The Lancet, 2018).

In Europe, CRC represents the second leading cancer type both in incidence and mortality considering both genders, whereas in Spain it is the first in incidence and only behind lung cancer regarding mortality. Each year 1 in 25 people will be diagnosed and 1 in 80 will die because of CRC in our country (Ferlay et al., 2019).

1.2 Etiology and risk factors

As a complex disease, CRC etiology involves a combination of both genetic and environmental factors. CRC familial risk is estimated to be around 12%-35%, according to twin studies (Lichtenstein et al., 2000; Jiao et al., 2014; Frank, Sundquist, Yu, Hemminki, & Hemminki, 2017), whereas the amount of hereditary CRC cases, linked to pathogenic genetic variants in high-risk cancer genes, is around 2 and 8% (Valle, de Voer, et al., 2019). Germline predisposition to CRC is further reviewed in chapter 2 of this thesis introduction.

Despite some non-modifiable risk factors, such as age or male gender, epidemiologic and migrant studies highlighted the importance of environmental risk factors in CRC etiology, according to the widely variation in incidence by world region

and also over time (Brenner & Chen, 2018; Murphy et al., 2019). Basically, Westernization of diet and lifestyle habits were found to be associated with an increasing incidence of CRC (Brenner et al., 2014). Among these potentially modifiable risk factors, tobacco smoking (Botteri et al., 2008), alcohol consumption (Bagnardi et al., 2015) and high intake of red and processed meats (Bouvard et al., 2015; Domingo & Nadal, 2017) are found among the most relevant. Overweight, obesity and physical inactivity were also considered as established causes for CRC (Boyle, Keegel, Bull, Heyworth, & Fritschi, 2012; Lauby-Secretan et al., 2016; Moore et al., 2016).

Inflammatory bowel disease is also a known risk factor for CRC, due to the associated chronic colitis. The risk increases with the duration of the disease (Jess, Rungoe, & Peyrin–Biroulet, 2012). On the other hand, protective effects were linked to regular intake of aspirin and other nonsteroidal anti-inflammatory drugs (Algra & Rothwell, 2012), as well as statins (Bardou, Barkun, & Martel, 2010) and hormone therapy in postmenopausal women (Limsui et al., 2012).

1.3 Molecular characterization

1.3.1 Precursor lesions and carcinogenesis pathways

CRC was one of the first solid tumors to be characterized at a molecular level, with some different signaling pathways implicated in the initiation and progress of the carcinogenesis (Fearon, 2011). This process was firstly described through the adenoma-carcinoma sequence by Vogelstein and collaborators, where an accumulation of genetic alterations both in oncogenes and tumor suppressor genes (TSGs) gives rise to the transition from a precursor lesion (called polyp or adenoma) to a carcinoma (**Figure 3**) (Vogelstein et al., 1988; Cho & Vogelstein, 1992). Oncogenes are defined as those genes leading to proteins whose activation is promoting tumorigenesis. Conversely, in the case of TSGs it is their loss of expression which is linked to the neoplastic phenotype development (Bashyam, Animireddy, Bala, Naz, & George, 2019).

Adenoma-carcinoma sequence starts with the formation of precursor lesions affecting isolated colonic crypts, known as aberrant crypt foci (Takayama et al., 1998). These evolve into early adenomas, presenting tubular histology, small size (< 1 cm) and typical intestinal type dysplastic morphology. The process continues through the advanced adenoma state, where the lesions acquire a villous component and a greater size (> 1 cm), before finally becoming a CRC (Kuipers et al., 2015). These different steps are characterized by specific genetic and/or epigenetic alterations that promote the acquisition of the neoplastic phenotype (**Figure 3**). This phenotype is mainly defined for an uncontrolled cell growth and the suppression of the cell death and repair mechanisms, as well as the acquisition of invasive and metastatic capacities. These and other key molecular features characteristic of the neoplastic progression in human

tumors were defined as the *hallmarks* of cancer by Hanahan and Weinberg in a seminal study (Hanahan & Weinberg, 2000) that was further improved with the latest advances in cancer research (**Figure 4**) (Hanahan & Weinberg, 2011). The alterations accumulated by the different precursor lesions can be different within the tumor, with some cancer cells acquiring specific mutations. According to the malignant potential of these mutations, this process leads to a clonal evolution, with some clones proliferating faster due to the acquired mutational events (Carethers & Jung, 2015).

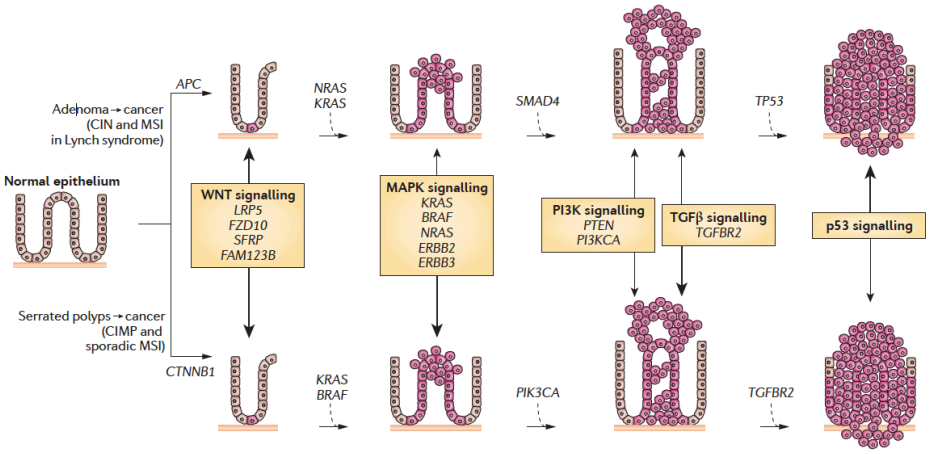


Figure 3. Adenoma-carcinoma sequence and serrated pathway of carcinogenesis. Colorectal cancer carcinogenesis pathways, precursor lesions and molecular alterations leading to the neoplastic development. CIMP, CpG island methylator phenotype; CIN, chromosomal instability; MSI, microsatellite instability (Kuipers et al., 2015).

Following this adenoma-carcinoma sequence, the first molecular defect for most colorectal tumors occurs in the *APC* gene, causing the deregulation of the Wnt/β-catenin signaling pathway (Kinzler & Vogelstein, 1996). In fact, *APC* mutations are found in over 70% of colorectal adenomas (Brenner et al., 2014). *APC* acts as a TSG in this pathway by regulating the levels of β-catenin and therefore some cellular mechanisms linked to the progress of the CRC phenotype, such as cell-cell adhesion, cell migration, chromosomal segregation and apoptosis (Fearon, 2011). Some other genetic and epigenetic alterations are accumulated over the neoplastic transformation, affecting key signaling pathways in cancer, including RAS–RAF–MAPK, PI3K–AKT, TGFβ and p53 pathways (Kuipers et al., 2015).

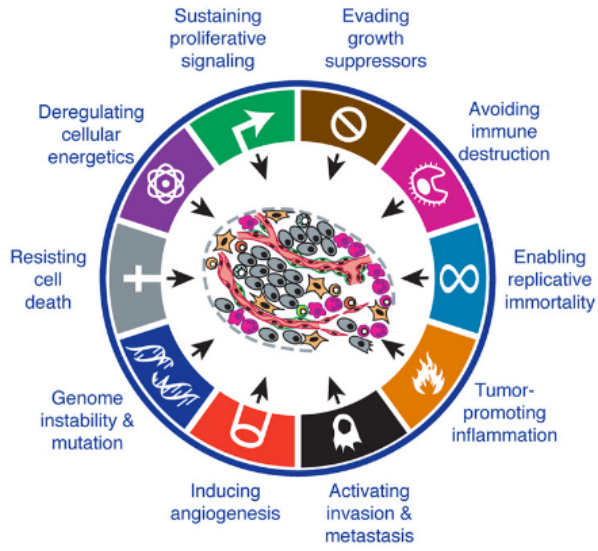


Figure 4. Hallmarks of cancer. Biological capabilities characterizing the development of human tumors (Hanahan & Weinberg, 2011).

In recent years, another different carcinogenesis pathway has been characterized, starting from a different precancerous lesion. This is the case of the serrated pathway, mainly characterized by the presence of a different type of polyps, called serrated polyps or lesions. These precursor lesions were considered indolent until the discovery of the serrated pathway, that is currently known to represent more than 15% of CRC cases (Carballal, Moreira, & Balaguer, 2013; IJspeert, Vermeulen, Meijer, & Dekker, 2015). Serrated lesions can be classified into five main categories according to the recent guidelines of the World Health Organization (WHO): sessile serrated lesions, sessile serrated lesions with dysplasia, traditional serrated adenomas, hyperplastic polyps and serrated adenomas unclassified (when there is no clear separation between traditional serrated adenomas and sessile serrated lesions) (Nagtegaal et al., 2019). They present histological and molecular features differentiated from traditional tubular adenomas (**Figure 3**). A common histological trait is a serrated architecture with invaginations of colonocytes in the lumen of the crypts, showing a stellate appearance in its cross section and a sawtooth shape in its longitudinal section. At a molecular level, serrated pathway is mainly characterized by the mutation of the *BRAF* oncogene, leading to the activation of the RAS-RAF-MAPK signaling pathway, although alternative *KRAS* mutations can also be found (Carballal et al., 2013; IJspeert et al., 2015). On the other hand, it is also common the inactivation of some TSGs via the hypermethylation of CpG islands on their promoting regions. This phenomenon is commonly known as the CpG island methylator phenotype (CIMP) (Toyota et al., 1999).

1.3.2 Molecular pathways

At a molecular level, three main pathways have been classically considered for colorectal carcinogenesis: chromosomal instability (CIN), microsatellite instability (MSI) and the already mentioned CIMP (**Figure 5**). However, they are not biologically exclusive, since tumors arising from precursor lesions following a particular pathway can accumulate genetic alterations typical of one of the others (Cunningham et al., 2010; Carethers & Jung, 2015). In fact, genetic instability *hallmark* is shared between CIN and MSI pathways, whereas MSI is also widely present in CIMP tumors. In addition, the biomarker defined by elevated microsatellite alterations at selected tetranucleotide repeats (EMAST) was established as a modulator for all three pathways and a potential predictor of patient survival (Devaraj et al., 2010; Carethers & Jung, 2015). This classic molecular classification is closely linked to the previously mentioned precursor lesions and their associated carcinogenic pathways. In this regard, CIN is associated to adenomas, whereas CIMP tumors are commonly originated by a serrated polyp. MSI can be found linked to both precursor lesions (Kuipers et al., 2015).

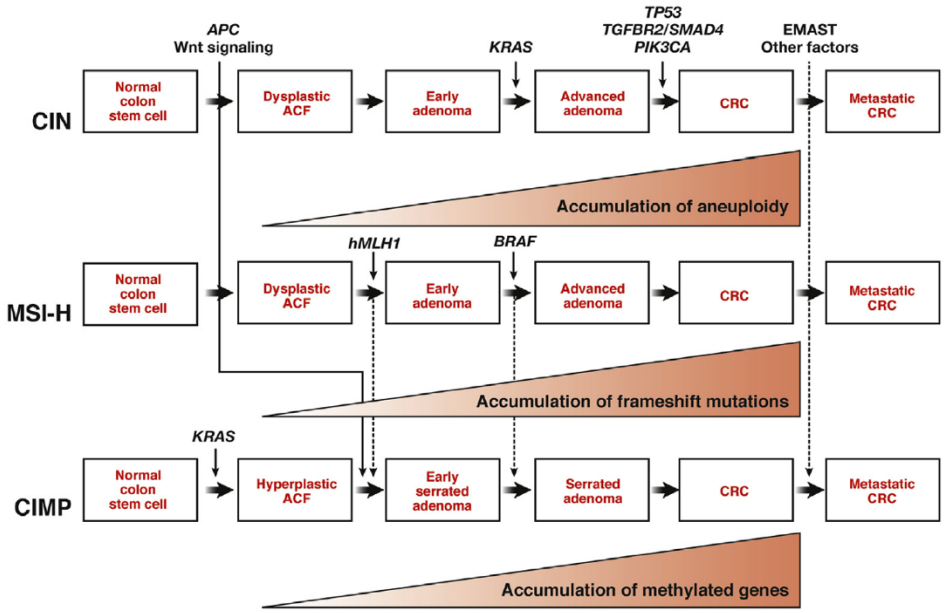


Figure 5. Molecular pathways of colorectal carcinogenesis. Three different pathways can lead to the development of a colorectal tumor. Each pathway is characterized by specific genetic and molecular alterations, with varying histology. CIMP, CpG island methylator phenotype; CIN, chromosomal instability; EMAST, elevated microsatellite alterations at selected tetranucleotide repeats; *hMLH1*, hypermethylated *MLH1*; MSI, microsatellite instability (Carethers & Jung, 2015).

CIN was the first molecular pathway described and is known to originate most cases of CRC, especially those from a sporadic background (up to 85% of sporadic CRCs) (Carethers & Jung, 2015). It is characterized by the accumulation of somatic copy number alterations (CNAs), including allelic losses of chromosomal arms 5q (where *APC* gene is located), 8p, 17p (*TP53*) and 18q (*SMAD4*), as well as gains of 8q, 13q and 20q (Ried et al., 1996; Hermsen et al., 2002; Cunningham et al., 2010; Fearon, 2011). All of these genetic events lead to tumor aneuploidy, defined as the state where the chromosome number in a cell is different from the normal state (which is diploid in the case of human cells, with a total of 46 chromosomes).

MSI is related to alterations in microsatellites, also called short tandem repeats and defined as stretches of repetitive DNA spread along the genome. These alterations appear in the form of small insertions or deletions (indels), leading to frameshift mutations and should be repaired by the DNA mismatch repair (MMR) system. However, when MMR is not functioning correctly the MSI phenotype appears (Grilley, Holmes, Yashar, & Modrich, 1990). MSI is therefore widely used as a biomarker for defective MMR in CRC (Ionov, Peinado, Malkhosyan, Shibata, & Perucho, 1993). This DNA repair deficiency is commonly caused by mutations in any of the MMR genes (*MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6* and *PMS2*), as well as somatic hypermethylation of *MLH1* promoter (Brenner et al., 2014; Carethers & Jung, 2015). This refers again to the connection of MSI with CIMP phenotype and serrated pathway, since *MLH1* could be one of the multiple methylation targets and is often silenced in tumors originated from serrated polyps (Carballal et al., 2013; IJspeert et al., 2015). MSI pathway is linked to a large accumulation of frameshift variants, potentially leading to disruption of protein translation if they are located in the coding region of a particular gene, as it is the case of driver genes *ACVR2*, *TGBR2*, *MSH3* and *MSH6*. On the other hand, MSI tumors present a very low number of chromosomal alterations in contrast to CIN tumors, thus maintaining a near diploid karyotype (Carethers & Jung, 2015). MSI is also closely linked to hypermutation, with approximately three-quarters of hypermutated CRCs found to harbor this phenotype among the cohort of *The Cancer Genome Atlas* (TCGA) project. However, interestingly most mutated samples of this study had somatic mutations in polymerase epsilon (encoded by *POLE*), but neither *MLH1* methylation nor MSI or CIMP phenotypes were present (Muzny et al., 2012).

CpG islands are regions of DNA enriched in this specific dinucleotide (CpG), observed in promoter and upstream regulatory regions of approximately 50% of the human genes. Hypermethylation of these areas of the genome suppresses the transcription of the affected gene, having a great importance in the case of TSGs for the carcinogenesis progress (Bird, 1986). The phenotype of CIMP is mainly defined according to the methylation status of a panel of 5 genes (*RUNX3*, *SOCS1*, *NEUROG1*, *CACNA1G* and *IGF2*), even though the criteria is not universally accepted. High CIMP is

considered when 3 or more of the 5 markers are hypermethylated, whereas low CIMP is present when they are 2 or less (Weisenberger et al., 2006). Interestingly, although CIMP phenotype is linked to hypermethylation of numerous cancer-associated genes, globally a generalized hypomethylation pattern exists across the whole genome of CRC cells in comparison with the adjacent normal tissue (Feinberg, Gehrke, Kuo, & Ehrlich, 1988). As previously commented, CIMP pathway is often linked to MSI phenotype, specially of sporadic origin linked to hypermethylation of *MLH1*. Those cases usually present gain-of-functions mutations in *BRAF* gene (particularly p.Val600Glu mutation) (Weisenberger et al., 2006; Goel & Boland, 2012). In addition, when CIMP is found in precursor lesions, it is commonly the case of serrated adenomas (Fearon, 2011; Dienstmann et al., 2017).

1.3.3 Consensus molecular subtypes

Recently, a new molecular classification has been described for CRC after the integration of all available biological data from previous CRC subtyping efforts. It is based on transcriptomic profiles, although data from mutations, CNAs, methylation, microRNAs and proteomics were also considered and integrated to obtained the so-called *Consensus Molecular Subtypes* (CMSs) (**Figure 6**) (Guinney et al., 2015).

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermutation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF-β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

Figure 6. Characterization of consensus molecular subtypes of colorectal cancer. Main molecular and cellular features displayed by the four transcriptomics-based molecular subtypes. CIMP, CpG island methylator phenotype; CMS, consensus molecular subtype; MSI, microsatellite instability; SCNA, somatic copy number alterations (Guinney et al., 2015).

The main achievement of the new molecular classification was to address a subclassification in the classic non-MSI/CIN group of CRC, accounting for up to 85% of the somatic cases. This subclassification considered three different groups, CMS2 (canonical), CMS3 (metabolic) and CMS4 (mesenchymal), with large differences in gene expression patterns. On the other hand, CMS1 accounted for almost all the cases following the classic MSI pathway, also linked with CIMP and hypermutation

phenotypes (only CMS3 presented some additional MSI samples).The availability of different layers of omics data, as well as clinicopathologic and prognostic records of the analyzed CRC samples, allowed to perform a complete characterization of the consensus gene expression-based subtypes, according to different molecular features, signaling pathway modulations and survival (**Figure 6**). This comprehensive characterization will facilitate the development of new personalized therapies for CRC treatment (Guinney et al., 2015; Dienstmann et al., 2017).

2. Germline predisposition to colorectal cancer

2.1 Techniques for the identification of new predisposition genes

2.1.1 Repertoire of genetic alterations

Regarding germline predisposition in complex diseases, as it is the case for CRC, the repertoire of genetic variants that can be affecting genes contributing to predisposition is very diverse. Variants can be classified according to their population frequency and their associated risk of developing the disease, known as penetrance (**Figure 7**) (McCarthy et al., 2008). High penetrance variants are defined as those causing a larger effect on the susceptibility to the disease, but also correspond to the rarest ones. Mainly linked to Mendelian disorders (Mendel, 1866), where the alteration of a single gene is often responsible for the phenotype, they have been classically identified by linkage studies. On the other hand, low penetrance variants, mainly identified by genome wide association studies (GWAS), are characterized by being common in the general population (generally with an allelic frequency over 5%) and having a little deleterious effect in disease development. A stronger effect would imply a decrease in the viability of carriers, which would not be compatible with a large allelic frequency. Despite of the small individual effect of this type of genetic alterations, a combination of them may contribute significantly to disease predisposition, along with the additive effect of environmental risk factors. This combination could be possible given their strong presence among the population (McCarthy et al., 2008; Manolio et al., 2009).

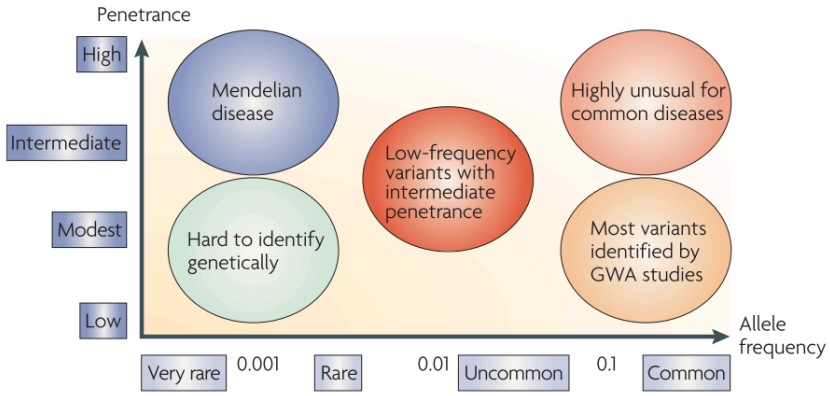


Figure 7. Classification of genetic variants according to germline allelic frequency and strength of deleterious potential in predisposition to a given disease. GWA, genome wide association (McCarthy et al., 2008).

In some conditions, as in the case of CRC, the ratio of estimated heritability according to classical family and twin studies (12-35%) is not in accordance with the heritability explained by the well-known genetic variants associated with the disease (2-

8%), therefore leading to what is known as *missing* heritability (Jiao et al., 2014; Valle, Vilar, Tavtigian, & Stoffel, 2019). Most of this *missing* heritability is hypothesized to remain in those variants not frequent enough to be identified by GWAS, nor having an effect size on disease development sufficient to be captured by classical linkage analysis in family studies (**Figure 7**) (Manolio et al., 2009).

2.1.2 Linkage analysis

Linkage disequilibrium is defined as the nonrandom association of alleles corresponding to two or more loci (Slatkin, 2008). This association is a consequence of the proximity and the corresponding low probability of recombination that would break the haplotype formed by the alleles (Collins, 2007). For the identification of genes implicated in predisposition, commonly informative markers such as microsatellites or single nucleotide polymorphisms (SNPs) are used. SNPs are defined as those genomic positions where two or more different nucleotides are common in the general population, i.e., variations with respect to reference genome that account for an elevated allelic frequency, often established as greater than 1% (Timpson, Greenwood, Soranzo, Lawson, & Richards, 2018). The association of a region containing some of these markers with the disease phenotype would be the basis for the identification of a putative predisposition gene. This region would be suspected to contain a highly penetrant variant in that gene, that would be inherited in a Mendelian fashion (Collins, 2007).

Accordingly, the identification of genes responsible for classic Mendelian CRC hereditary syndromes was allowed by linkage studies, including *APC* (Bodmer et al., 1987; Leppert et al., 1987), *MLH1* (Lindblom, Tannergård, Werelius, & Nordenskjöld, 1993) and *MSH2* (Peltomaki et al., 1993) among others (**Figure 8**). In the case of the detection of variants with a medium effect in disease predisposition, linkage analysis presents a lower power and resolution, thus limiting its success (Manolio et al., 2009). However, in last years, in combination with novel sequencing techniques, linkage studies have turned into a promising approach for solving the missing heritability linked with medium penetrance variants in complex traits (Ott, Wang, & Leal, 2015). Some recent successful examples have been published for different diseases (Norton et al., 2013; Eggers et al., 2015; Toma et al., 2018), also including CRC after a study conducted in collaboration with our research group (Toma et al., 2019).

2.1.3 Genome wide association studies

GWAS studies are based on the association of complex traits with common variants in the form of SNPs, therefore having an individual moderate effect on disease susceptibility (i.e. low penetrance variants). In this regard, SNP genotyping is performed in large cohorts of cases and controls along the entire human genome, thus leading to

an agnostic approach regarding identification of new genes involved in a certain disease. However, even if substantial contributions were made by this technology, deciphering of the molecular mechanisms behind the identified associations has been challenging, thus limiting potential applicability of the results (Ott et al., 2015; Sud, Kinnersley, & Houlston, 2017).

Regarding CRC, GWAS studies have been conducted since 2007, allowing the identification of around 130 common-low penetrance variants that could account for up to 7-8% of the susceptibility to this disease, considering their additive effect and also the combination with the environmental risk factors (**Figure 8**) (Jiao et al., 2014; Peters, Bien, & Zubair, 2015; Buniello et al., 2019).

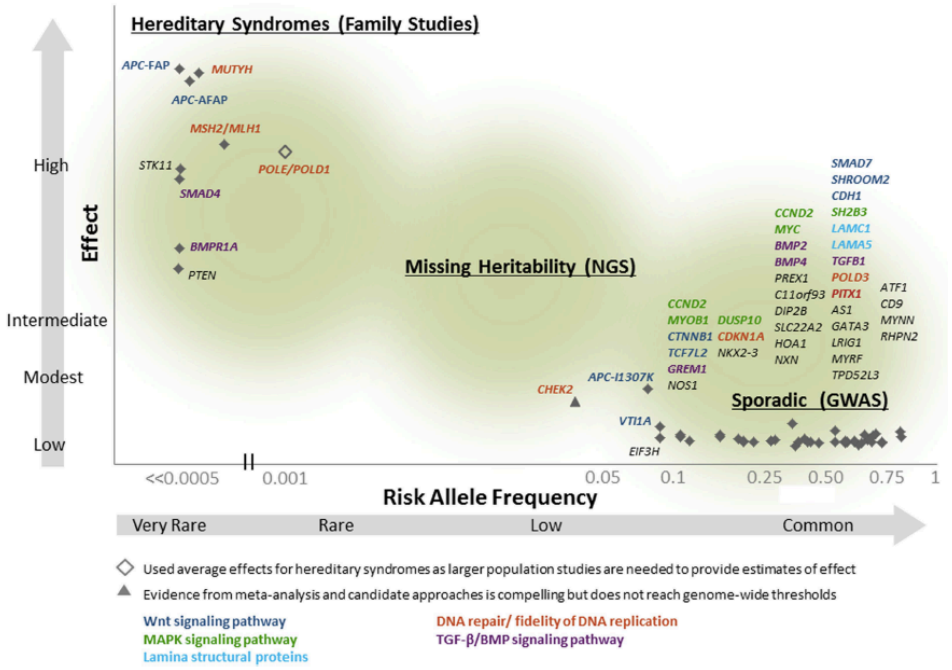


Figure 8. Spectrum of colorectal cancer susceptibility genes. Classification of known genes involved in colorectal cancer predisposition according to allele frequency and penetrance, as well as their associated molecular pathway. AFAP, attenuated familial adenomatous polyposis; FAP, familial adenomatous polyposis; GWAS, genome wide association studies; NGS, next generation sequencing (Peters et al., 2015).

2.1.4 Next generation sequencing

Currently, the most used method for the identification of new genes implicated in the predisposition to a certain disease is next generation sequencing (NGS). This technique has led to a revolution in the genetics field, allowing the simultaneous sequencing of millions of short DNA fragments (called *reads*) that generates a large volume of information at a very reduced cost when comparing to previous technologies

(Lappalainen, Scott, Brandt, & Hall, 2019). However, handling of the considerable amount of data produced by NGS has forced to create multidisciplinary analysis teams, due to the resulting statistical and bioinformatic burden (Metzker, 2010; Goodwin, McPherson, & McCombie, 2016). NGS tackles the commented missing heritability gap, since it allows to focus on those rare variants with high or moderate effect on the disease development that cannot be identified by linkage or GWAS (**Figure 8**) (Manolio et al., 2009; Peters et al., 2015).

NGS applications include whole genome sequencing (WGS), although in the case of translational biomedicine the sequencing directed to the coding regions of the genome (i.e. exons), also known as whole exome sequencing (WES), has become the most successful approach. Capture of exon regions facilitated the investigation on a crucial part of the human genome with the advantage of saving costs with respect to WGS, thus allowing an increase in the number of samples sequenced. This sequencing technology also permits testing specific gene panels, although this is more frequently used for diagnosis in the clinical setting (basically for financial reasons). However, with respect to gene panels, WES enables the unbiased approach needed for the identification of new genes linked to a certain disease, rather than limiting the study to what is already known. Information of all coding genes is available, therefore no previous assumption is needed about the genes or pathways implicated in a specific studied phenotype (Teer & Mullikin, 2010; Goodwin et al., 2016).

NGS has been commonly used for single nucleotide variants (SNVs) and indels, genetic alterations that affect a reduced number of nucleotides (even just one base substituted in the case of SNVs). Considering the usual length of sequencing reads (around hundreds of base pairs for mainly used sequencing technologies (Loman et al., 2012; Goodwin et al., 2016)), both types of alterations are invariably contained inside a unique read, increasing accuracy. On the other hand, NGS has also marked a turning point regarding copy number variants (CNVs) identification. This variant typology is defined as those DNA fragments with a size over 50 base pairs with variations in copy number (deletions or duplications) in comparison with the human reference genome, thus generating an alteration in the basal diploid status. Therefore, CNVs are also described as unbalanced alterations, forming together with the so-called balanced changes (mainly inversions and translocations) and different types of insertions (from novel sequences or mobile elements) the whole spectrum of structural variants (SVs) (Alkan, Coe, & Eichler, 2011; MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014; Tattini, D'Aurizio, & Magi, 2015). Along with indels, CNVs represent more than 10 times more variation in the human genome than SNVs (Pang et al., 2010). CNVs have also been implicated in germline predisposition to different diseases, including CRC, where different genes have been highlighted after CNV profiling efforts, such as *EPCAM* (Ligtenberg et al., 2009), *GREM1* (Jaeger et al., 2012), *BUB1* (de Voer et al., 2013), *FOCAD*

(Weren, Venkatachalam, et al., 2015) or *TMEM158* (Franch-Expósito et al., 2018). They exert their influence mainly by modifying the expression of genes contained in the rearranged region, even though an indirect way is also possible by affecting downstream signaling pathways or regulatory regions (Henrichsen, Chaignat, & Reymond, 2009). Different bioinformatic tools and algorithms have been developed during last years in order to identify CNVs from NGS data, especially for WGS, overcoming the potential issue of CNVs spanning through lots of different sequencing reads. Most approaches can also be adapted for WES, although the intrinsic sparseness of read depth data of this technology makes the calling of CNVs particularly challenging (Kadalayil et al., 2015; Tattini et al., 2015).

In order to find new genes involved in germline predisposition to a particular disease, NGS permits the identification of a great number of genetic variants in every patient, thus emerging the need of a prioritization strategy (Ott et al., 2015). Different strategies can be explored in this regard, including the sequencing of different members of the same family, allowing for example to discard those alterations not shared among all affected relatives, as well as the use of large population germline variation databases such as Exome Aggregation Consortium (ExAC) (Lek et al., 2016) or Genome Aggregation Database (gnomAD) (Karczewski et al., 2019) (**Figure 9**). Additionally, as recently recommended by the Clinical Genome Resource, somatic mutation profiling can also be used in this regard, in order to help in the identification of the putative germline deficiencies responsible for the tumor phenotype encountered (Walsh et al., 2018).

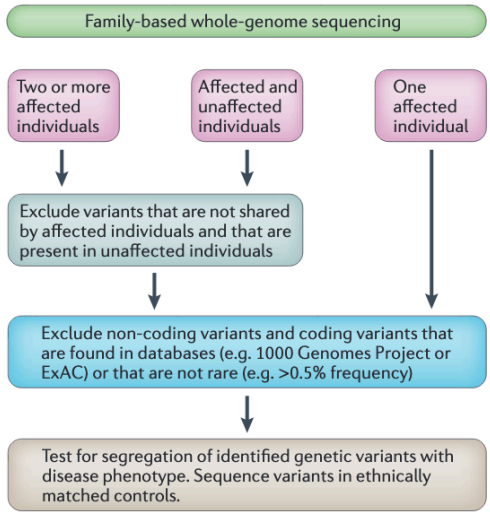


Figure 9. Filtering workflow for family-based next generation sequencing. Different filtering steps available according to the study design selected for the identification of novel genes implicated in disease predisposition. ExAC, exome aggregation consortium (Ott et al., 2015).

2.2 Hereditary syndromes

2.2.1 Overview

Inherited predisposition syndromes to CRC related to high penetrance genetic variants are behind 2-8% of all cases, and up to 6-10% if moderate penetrance variants are also considered (Valle, Vilar, et al., 2019). Different genes, belonging to different molecular pathways are affected, giving rise to a spectrum of hereditary syndromes. These syndromes are characterized by an increased risk of CRC compared to normal population, as well as for being originated from different types of preneoplastic lesions (i.e. polyps) (**Figure 10**) (Tomlinson, 2015). Accordingly, they can be phenotypically classified in polyposis and non-polyposis CRC syndromes, based on the presence or not of an accumulation of multiple preneoplastic lesions (**Figure 11**).

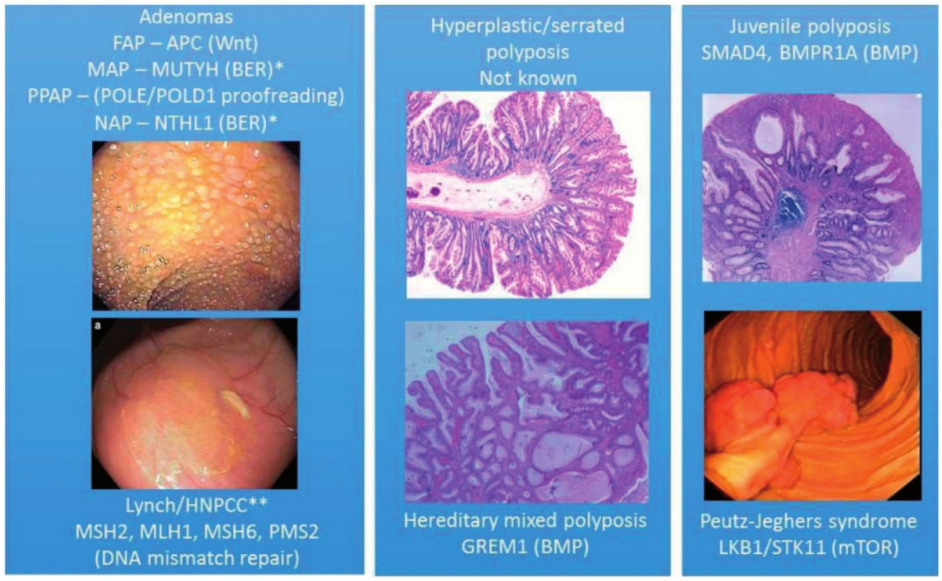


Figure 10. Repertoire of Mendelian colorectal cancer syndromes. Histological and colonoscopy images of the different colorectal cancer hereditary syndromes (most of them associated to an autosomal dominant inheritance pattern), linked to each corresponding predisposition lesion and causal gene. *, recessive inheritance; **, also mutated in the recessive syndrome congenital mismatch repair deficiency, where predisposition lesions are commonly conventional adenomas; BER, base excision repair; FAP, familial adenomatous polyposis; HNPCC, hereditary non-polyposis colorectal cancer, MAP, *MUTYH*-associated polyposis; NAP, *NTHL1*-associated polyposis; PPAP, polymerase proofreading-associated polyposis (Tomlinson, 2015).

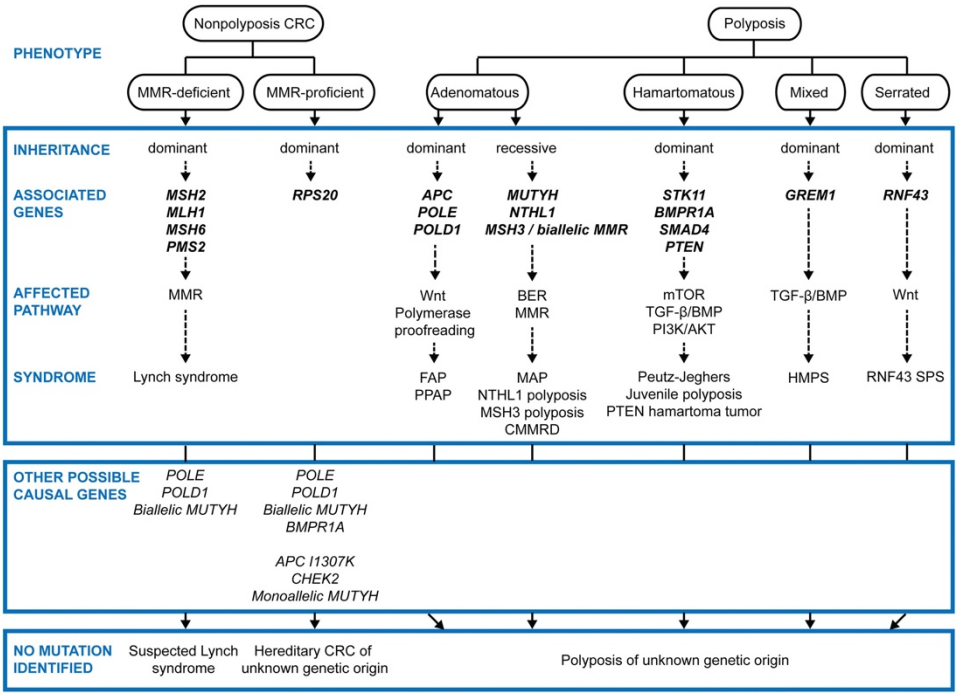


Figure 11. Classification of hereditary CRC syndromes based on predisposition lesions. Spectrum of syndromes including mode of inheritance, gene deficiencies and affected pathways associated. BER, base excision repair; CMMRD, constitutional mismatch repair deficiency; CRC, colorectal cancer; HMPS, hereditary mixed polyposis syndrome; MAP, *MUTYH*-associated polyposis; MMR, DNA mismatch repair; PPAP, polymerase proofreading-associated polyposis; SPS, serrated polyposis syndrome (Valle, Vilar, et al., 2019).

2.2.2 Lynch syndrome

Previously known as hereditary non-polyposis CRC, Lynch syndrome constitutes one of the first hereditary cancer syndromes described (H. T. Lynch & Krush, 1971), as well as the most frequent predisposition syndrome to CRC, accounting for up to 4% of total CRC patients and a lifetime risk of 50-80% (Jasperson, Tuohy, Neklason, & Burt, 2010; Yurgelun et al., 2017). It is characterized by pathogenic germline mutations in four of the DNA MMR genes, namely *MLH1* (Lindblom et al., 1993; Bronner et al., 1994; Papadopoulos et al., 1994), *MSH2* (Fishel et al., 1993; Leach et al., 1993; Peltomaki et al., 1993), *MSH6* (Miyaki et al., 1997) and *PMS2* (Nicolaides et al., 1994), with an autosomal dominant pattern of inheritance and a primary predisposition to carcinomas, frequently with a low number of polyps (Moreira et al., 2012; Tomlinson, 2015). Interestingly, deletions on the 3' region of *EPCAM* gene (located upstream of *MSH2*) were also identified as a cause of Lynch syndrome (Ligtenberg et al., 2009). Apart from CRC, Lynch syndrome patients have an important risk of developing endometrial cancer (40-60%), as well as an increased cancer risk compared to the general population for gastric,

ovarian, small bowel, urinary tract, brain, pancreatic, prostate and skin (sebaceous tumors) cancers (Yurgelun et al., 2017). Lynch syndrome tumors typically present MSI phenotype, as well as loss of expression of the corresponding DNA MMR protein, commonly detected by immunohistochemistry. Indeed, as no germline genetic analysis are frequently performed in all CRC patients for Lynch syndrome diagnosis, these two molecular alterations are the foundation for suitable patient selection in clinical settings. In this regard, different clinical criteria for the identification of the families most likely to carry this syndrome were developed, in order to avoid universal molecular characterization (H. T. Lynch, Snyder, Shaw, Heinen, & Hitchins, 2015). Expected familial aggregation for CRC was considered for the initial development of these criteria, known as Amsterdam I criteria (Vasen, Mecklin, Khan, & Lynch, 1991), while extracolonic neoplasms were added in subsequent renewed versions, Amsterdam II criteria (Vasen, Watson, Mecklin, & Lynch, 1999) and revised Bethesda guidelines (Umar et al., 2004) (**Figure 12**).

Amsterdam II criteria for Lynch Syndrome	Revised Bethesda guidelines for MSI testing in CRC
1. Three relatives diagnosed with a LS-associated cancer (with one being a first-degree relative of the other two)	1. CRC diagnosed in a patient under the age of 50 years
2. At least two successive generations affected	2. Presence of synchronous or metachronous LS-associated cancer, regardless of age
3. At least one of the patients of a LS-associated cancer diagnosed before 50 years of age	3. CRC with MSI-high histology diagnosed in a patient under the age of 60 years
4. Exclusion of FAP in CRC cases	4. CRC diagnosed in a patient with at least one first-degree relative with a LS-associated cancer, with one of the cancers diagnosed before the age of 50 years
5. Verification of tumors by pathology whenever possible	5. CRC diagnosed in a patient with at least two first- or second-degree relatives with a LS-associated cancer, regardless of age

Figure 12. Standardized clinical guidelines for Lynch syndrome diagnosis. Amsterdam II criteria for Lynch syndrome diagnosis and revised Bethesda guidelines for the selection of those Lynch syndrome patients suitable for microsatellite instability testing. Note: Lynch syndrome-associated cancers include endometrium, stomach, ovary, pancreas, ureter or renal pelvis, brain, small bowel, hepatobiliary tract and skin (sebaceous tumors). CRC, colorectal cancer; FAP, familial adenomatous polyposis; LS, Lynch syndrome; MSI, microsatellite instability. Adapted from H. T. Lynch et al., 2015.

However, approximately half of the CRC families fulfilling Amsterdam criteria present MMR proficient tumors without alterations of the MMR system. This phenotype, classically known as familial CRC type X (Lindor et al., 2005), presents a lower lifetime risk of CRC than Lynch syndrome, no increased risk of extracolonic

neoplasias and a diagnosis of CRC around 10 years later on average (Jasperson et al., 2010; Peters et al., 2015). Even though a huge effort has been made in order to found potential causal genes for MMR proficient hereditary non-polyposis CRC, only *RPS20*, encoding for a component of the small ribosomal subunit, was found in a study combining linkage analysis and WES with consistent evidence, suggesting a high penetrance although additional cases are needed for better risk estimations (Nieminen et al., 2014).

In those rare cases when two germline pathogenic alterations are detected in the DNA MMR genes associated to Lynch syndrome, constitutional mismatch repair deficiency (CMMRD) syndrome is diagnosed (Ricciardone et al., 1999; Wang et al., 1999). A more severe phenotype is found in this syndrome, with a high risk of developing a spectrum of different neoplasias at a young age, including T-cell non-Hodgkin lymphomas, high-grade gliomas and gastrointestinal, mainly colorectal, cancers, as well as some typical features of neurofibromatosis patients, namely café au lait spots (Bakry et al., 2014). Interestingly, adenomatous polyposis has been recently reported in patients harboring biallelic inactivation of other two MMR genes, *MSH3* (Adam et al., 2016) and *MLH3* (Olkinuora et al., 2019), as well as in CMMRD patients (Aronson et al., 2016).

2.2.3 Familial adenomatous polyposis

Second most common hereditary CRC syndrome is familial adenomatous polyposis (FAP), responsible for up to 1% of all CRCs (Kanth, Grimmett, Champine, Burt, & Samadder, 2017). An autosomal dominant pattern of inheritance is also showed by FAP patients, as well as germline genetic defects in *APC* gene (Bodmer et al., 1987; Leppert et al., 1987). Risk of CRC in those patients harboring *APC* mutations is up to 100%, therefore screening is recommended to be initiated at a young age (around 10-15 years of age, according to the US National Comprehensive Cancer Network) and surgical colectomy is often required (Kanth et al., 2017; Gupta et al., 2019). Clinical phenotype is in most cases defined by adenomatous polyposis, i.e. a great accumulation of conventional adenomas throughout the colon and rectum, from hundreds to thousands (Bussey, 1975). However, milder phenotypes, where the presence of colorectal polyps is reduced (normally between 10 and 100) are also found and typically referred to as attenuated FAP (H. T. Lynch et al., 1988, 1995; Leppert et al., 1990). In this case, autosomal dominant pattern of inheritance is also present and predisposition is also driven by *APC* germline deficiency, although with specifically located mutations, including the extreme 5' (exons 3-4 and intron 3), exon 6, exon 9, intron 9 and the 3' end of the gene (exon 15) (Church, Hernegger, Moore, & Guillem, 2002), as well as the deletion of the entire *APC* gene (Pilarski, Brothman, Benn, & Shulman Rosengren, 1999). Germline SNVs in *APC* promoter 1B have also been recently found to cause gastric

adenocarcinoma and proximal polyposis of the stomach (GAPPS) syndrome. Interestingly, GAPPS patients present a large number of gastric polyps (fundic gland type), as well as a high risk of gastric cancer, but not CRC or colorectal polyposis (Li et al., 2016).

2.2.4 *MUTYH*-associated polyposis

MUTYH-associated polyposis (MAP) is an autosomal recessive inherited condition characterized by biallelic germline alterations in the base excision repair (BER) gene *MUTYH*, a lifetime risk of CRC of around 80% and the presence of classic or attenuated adenomatous polyposis. However, a higher prevalence of serrated polyps was observed for MAP in comparison with FAP and attenuated FAP syndromes (Al-Tassan et al., 2002; Kanth et al., 2017). Monoallelic variants in *MUTYH* have also shown a moderate increase in CRC risk, particularly in those cases with a first-degree relative diagnosed with CRC (Win et al., 2014). Interestingly, a particular somatic mutational profile was found in tumors of MAP patients, with predominance of G:C>T:A SNVs, that will be further discussed in section 3 of this introduction (Pilati et al., 2017; Viel et al., 2017).

2.2.5 Polymerase proofreading-associated polyposis

Recently, replicative and repair DNA polymerases *POLE* and *POLD1* have been identified using NGS (particularly WGS) to cause hereditary predisposition to CRC and adenomatous polyposis following an autosomal dominant pattern (Palles et al., 2013). Missense pathogenic mutations in the exonuclease domain of both genes, in charge of polymerase proofreading DNA repair activity, lead to the so-called polymerase proofreading-associated polyposis (PPAP) syndrome. Phenotype of PPAP patients is mainly characterized by the presence of multiple adenomas, as well as an increased risk of CRC and also of endometrial cancer in the case of female carriers of pathogenic variants in *POLD1* (Bellido et al., 2016). Spectrum of PPAP-associated cancers has been recently expanded, including ovarian, brain (Rohlin et al., 2014), pancreatic and small bowel cancers (Hansen et al., 2015), melanoma (Aoude et al., 2015) and a clinical phenotype suggestive of CMMRD (Wimmer et al., 2017). Tumors harboring deleterious germline mutations in *POLE* and *POLD1* frequently present somatic hypermutation, along with particular mutational profiles for each of the genes and usually a MMR proficient phenotype (Muzny et al., 2012; Alexandrov et al., 2019). However, interestingly concomitant *POLE* and *POLD1* germline mutations and somatic alterations in MMR genes have also been recently found in suspected Lynch syndrome cases (i.e. those patients harboring MMR deficiency but without germline alterations in known Lynch syndrome-associated genes), suggesting that the somatic inactivation of the MMR system is a consequence of the hypermutator phenotype linked to germline alterations in *POLE*/*POLD1* (Elsayed et al., 2015; Jansen et al., 2016).

2.2.6 *NTHL1*-associated tumor syndrome

NTHL1 biallelic germline pathogenic variants were found linked to predisposition to multiple adenomas and CRC in a study using WES in 51 patients (Weren, Ligtenberg, et al., 2015). However, the scope of tumors associated to this BER gene has been recently broadened, including 14 different cancer types, thus leading to the extensive denomination of *NTHL1*-associated tumor syndrome (NATS) (Grolleman, de Voer, et al., 2019). NATS is linked to an autosomal recessive pattern of inheritance, with initial cases harboring pathogenic homozygous variants but also with some recent cases found carrying compound heterozygous mutations. In comparison with *MUTYH*, even if both genes belong to the same signaling pathway, *NTHL1* deficiency has been linked to an increased risk to a wider repertoire of cancer types, although being at least five times less prevalent than MAP (Valle, de Voer, et al., 2019). Interestingly, *NTHL1* was validated among a full set of novel genes proposed as candidates por germline predisposition to CRC (Broderick et al., 2017). As in the case of the other CRC predisposition syndromes arising from defects in DNA repair-associated genes, NATS-associated tumors harbor a specific mutational profile. In this case, it is characterized by a predominance of C>T transitions at non-CpG sites (Weren, Ligtenberg, et al., 2015; Grolleman, de Voer, et al., 2019), and has been recently validated using human intestinal organoids (Drost et al., 2017).

2.2.7 Serrated polyposis syndrome

Serrated polyposis syndrome (SPS) is a clinical condition of recent diagnosis characterized by the colonic presence of multiple and/or large serrated polyps, as well as a moderate risk of CRC (around 16% according to latest studies). SPS polyps present particular features differentiating them from sporadic serrated polyps. They are often located in proximal colon and they are presented frequently in an abnormal size and number. It is also possible to find serrated polyps of small size but spread along the colon and rectum (Carballal et al., 2016; IJspeert et al., 2015). SPS prevalence is unknown, but some studies reported a low value for CRC screening programs using primary colonoscopy (<0.1%). A higher value was found for those screening populations where a fecal occult blood test was used for improved patient selection previously to the colonoscopy (0.34%-0.66%) (Biswas et al., 2013; Moreira et al., 2013; Carballal et al., 2016). These values are much higher than initially though, which could be linked to the difficult endoscopic detection of serrated polyps, due to its usual proximal location, flat morphology and similar coloring than surrounding mucosa. Additionally, consensus was also challenging for the anatomopathological classification of these polyps (Carballal et al., 2013). Considering these discrepancies, in 2010 the WHO defined the following diagnostic criteria for SPS: i) at least 5 serrated polyps proximal to the sigmoid colon, two or more larger than 10 mm; ii) any number of serrated polyps proximal to the

sigmoid colon in an individual with one first-degree relative diagnosed with SPS; or iii) more than 20 serrated polyps of any size distributed throughout the colon (Snover, Ahnen, Burt, & Odze, 2010). However, these criteria have been recently updated in the new classification of the tumors of the digestive system by the WHO in the current year 2019. Thus, previous criterion ii was discarded according to the lack of evidence proven along the years, while the other two were reformulated as follows: i) at least 5 serrated polyps proximal to the rectum, all larger than 5 mm and two or more larger than 10 mm; or ii) more than 20 serrated polyps of any size distributed throughout the large bowel, with at least 5 being proximal to the rectum. These changes were mainly based on the fact that around 50% of CRCs in SPS patients are found in the recto-sigmoid (Nagtegaal et al., 2019).

Regarding predisposition, it has been argued that SPS does not constitute an inherited genetic syndrome, due to the usual late age of onset (between 50 and 60 years) and the strong association with environmental factors, namely alcohol consumption, tobacco smoking and fat intake (Buchanan et al., 2010; Jasperson et al., 2013; IJspeert et al., 2017). However, a candidate predisposition gene was suggested by some recent studies, although with some controversy, the inhibitor of the Wnt/β-catenin signaling pathway *RNF43*. Pathogenic germline mutations were found in a total of 12 patients of SPS and/or CRC belonging to 7 different families, with 50% of the colonic lesions analyzed showing the serrated pathway characteristic CIMP (Gala et al., 2014; Taupin et al., 2015; Buchanan et al., 2017; Yan et al., 2017; Quintana et al., 2018).

2.2.8 Other predisposition syndromes

Additional preneoplastic lesions apart from conventional adenomas and serrated lesions, known as hamartomas, give rise to a range of autosomal dominant hereditary CRC syndromes. Hamartomas are histologically characterized by a tree-like configuration with arborizing strands of smooth muscle and dilated crypts (H. Ma et al., 2018). The so-called hamartomatous polyposis syndromes are rare, with a prevalence ten times lower than previously mentioned adenomatous polyposis syndromes. They have been associated with pathogenic germline variants in genes such as *STK11* (Peutz-Jeghers syndrome) (Giardiello et al., 1987), *BMPR1A* (Howe et al., 2001), *SMAD4* (Howe et al., 1998) (juvenile polyposis syndrome) and *PTEN* (*PTEN*-hamartoma tumor syndrome / Cowden syndrome) (Liaw et al., 1997). Additionally, hereditary mixed polyposis syndrome is defined by the accumulation of multiple colorectal polyps of different histology types, including conventional adenomas, serrated lesions and hamartomas, leading to an increased risk to CRC. Duplications affecting the upstream regulatory region of *GREM1* gene have been recently linked to this inherited syndrome (Jaeger et al., 2012; Rohlin et al., 2016).

2.3 Familial colorectal cancer

Apart from the commented inherited predisposition syndromes, linked to genetic alterations specifically affecting some particular well-known genes, genetic factors are expected to be responsible for 12-35% of CRC cases (Lichtenstein et al., 2000; Jiao et al., 2014; Peters et al., 2015). As only up to 8% of CRCs are explained by known high penetrance variants (Yurgelun et al., 2017; Valle, Vilar, et al., 2019), a missing heritability is present and has been targeted by different studies looking for new potential candidate genes that can have an substantial impact on genetic counseling in the affected families. NGS has been used as the main technology in this candidate gene identification effort for CRC predisposition (Valle, 2017; Valle, de Voer, et al., 2019). A large number of studies have been published in last years, resulting in a long list of genes proposed to be involved in hereditary CRC, including *BUB1*, *BUB3*, *SEMA4A*, *FAN1*, *BLM*, *MCM9*, *FOCAD*, *MIA3*, *SETD6* and *BRF1* among the most promising candidates.

BUB1 and *BUB3* were found germline mutated in a cohort of 62 patients of early onset CRC, using copy number profiling by SNP arrays and multiplex ligation-dependent probe amplification (MLPA), as well as WES for the variant identification. Implicated in the maintenance of chromosomal stability, functional studies in human CRC cell lines revealed the potential role of this gene family in CRC predisposition (de Voer et al., 2013). Results were replicated in an additional cohort of 146 familial or early-onset CRC patients within the same study, as well as in an independent cohort of 456 MMR proficient hereditary non-polyposis CRC cases and 88 polyposis cases. However, low frequency found for variants in both *BUB1* and *BUB3*, as well as the lack of functional effects of some of the initially identified do not support the need to include these genes in routine germline genetic testing for CRC predisposition (Mur et al., 2018).

SEMA4A, encoding for a semaphoring protein, was proposed as candidate for predisposition to familial CRC type X, after the identification of a germline missense mutation in an Austrian family. Functional effects generated by this gene deficiency corroborated its putative implication in hereditary CRC (Schulz et al., 2014). Subsequently, validation in an additional cohort of early onset/familial CRC was unsuccessful for this gene (Kinnersley et al., 2016). However, in a re-analysis of the latter study, the original authors provide additional evidence to conclude that *SEMA4A* could remain as a susceptible candidate gene specifically for the familial CRC type X phenotype (Sill, Schulz, Steinke-Lange, & Boland, 2016).

FAN1 corresponds to a DNA repair gene that was initially proposed as the causal gene in a familial CRC type X family carrying a truncating germline mutation. Validation was also performed by *in vitro* functional studies and replication in a cohort of 176 additional families with familial CRC type X (Seguí et al., 2015). Interestingly, *FAN1* is linked to the Fanconi anemia pathway, which has been recently found involved in CRC

predisposition by our research group (Esteban-Jurado et al., 2016). Conversely, association of *FAN1* with CRC predisposition was discarded by two subsequent studies (Broderick et al., 2017; Fievet et al., 2019).

Biallelic germline mutations in RecQ helicase *BLM* have been classically associated with cancer hereditary syndrome Bloom syndrome (Ellis et al., 1995). Likewise, monoallelic mutations in this gene were proposed to increased CRC risk (Gruber et al., 2002). Subsequently, this gene was associated with predisposition both in breast (Thompson et al., 2012) and CRCs (de Voer et al., 2015). In this latter case, germline heterozygous variants in *BLM* were found enriched in early onset CRC cases compared to controls, after the initial detection by WES in a cohort of 55 early onset CRC patients (de Voer et al., 2015).

Another DNA helicase, *MCM9*, implicated in double-strand break repair via homologous recombination, was recently proposed as candidate for CRC predisposition after the detection of a homozygous frameshift mutation in a family with hereditary mixed polyposis, early onset CRC and primary ovarian failure (Goldberg et al., 2015). Interestingly, biallelic mutations in this gene had been linked to primary ovarian failure in a previous study (Wood-Trageser et al., 2014).

FOCAD encodes for a focal adhesion protein proposed to function as tumor suppressor in gliomas (Brockschmidt et al., 2012). However, this gene was also found implicated in polyposis and CRC predisposition after the identification of a germline intragenic deletion by performing CNV analysis (Weren, Venkatachalam, et al., 2015). The assessment of this kind of variants has also allowed to propose a rare duplication event as the underlying cause of the germline predisposition to familial CRC in a recent study of our research group using WES (Franch-Expósito et al., 2018).

Recently, three additional genes have been suggested as candidates for familial CRC predisposition, *MIA3*, *SETD6* and *BRF1*. Using homozygosity mapping in a cohort of 302 CRC cases and 3,367 controls, and subsequent linkage analysis, WES and WGS in a particular family with microsatellite stable CRC *MIA3* was pinpointed as the causal predisposition gene (Schubert et al., 2017). Likewise, germline alterations in *SETD6* and *BRF1* have been recently linked to hereditary CRC after functional validation in human CRC cell lines and yeast (Martín-Morales et al., 2017; Bellido et al., 2018).

Additional candidate genes were proposed by different studies, although with less evidence to be implicated in hereditary CRC, including *PTPRJ* (Venkatachalam et al., 2010; Hansen et al., 2017), *GALNT12* (Guda et al., 2009; Seguí et al., 2014; Evans et al., 2018), *FANCM*, *LAMB4*, *LAMC3*, *NOTCH3*, *PTCHD3*, *TREX2* (C. G. Smith et al., 2013), *CENPE*, *KIF23* (DeRycke et al., 2013), *AKR1C4*, *CCDC18*, *MRPL3*, *NUDT7*, *PRADC1*, *PRSS37*, *PSPH*, *SFXN4*, *TWSG1*, *UACA*,*ZNF490* (Gylfe et al., 2013), BARD1, *CDKN1B*, *EPHX1*, *NFKBIZ*, *SMARCA4*, *XRCC4* (Esteban-Jurado et al., 2015), *SMAD9* (Ngeow et al.,

2015), *ERCC6*, *WRN* (Arora et al., 2015), *LRP6*, *PTPN12* (de Voer et al., 2016), *DSC2*, *PIEZO1*, *ZSWIM7* (Spier et al., 2016), *MRE11*, *POLE2*, *POT1* (Chubb, Broderick, Dobbins, Frampton, et al., 2016), *BRCA2*, *BRIP1*, *FANCC*, *FANCE*, *REV3L* (Esteban-Jurado et al., 2016), *AK3*, *SLIT2* (Brea-Fernandez et al., 2017), *AXIN1*, *BMP4*, *CCDC18*, *NUDT7*, *PICALM*, *SLC5A9*, *TLR2*, *TWSG1*, *UBAP2*, *USP6NL*, *ZFP14* (Hansen et al., 2017), *CEBPZ*, *DDX20*, *ETAA1*, *FAT1*, *IFRD2*, *LRBA*, *PIK3R3*, *SEMA3G*, *SLC11A1*, *SLC26A8*, *TP53INP1*, *ZEB2*, *ZFYVE26* (L. Yu et al., 2018), *TMEM158* (Franch-Expósito et al., 2018) and *MGMT* (Belhadj et al., 2019).

However, no functional studies linking the proposed candidates with the molecular pathogenesis of familial CRC have been performed in most cases, thus lacking a robust evidence of causality (Valle, de Voer, et al., 2019). In fact, a recent study based on a cohort of 863 familial CRC cases (defined as having an age of diagnosis below 55 years and at least one first-degree relative with CRC) and 1,604 controls found insufficient the evidence to claim as CRC predisposition genes some of the strongest candidates to date. Only the already mentioned *NTHL1* and *RPS20* genes were found to accumulate enough evidence as hereditary CRC genes. In this regard, segregation of the genotype with the phenotype in families, somatic mutations and functional studies were considered, as well as a case-control analysis based on the provided familial CRC cohort (Broderick et al., 2017).

3. Somatic mutational profiling

3.1 Knudson's two-hit hypothesis

The idea that cancer arises from the accumulation of multiple genetic alterations was firstly developed by Carl O. Nordling in 1953 (Nordling, 1953). Taking this idea, Alfred G. Knudson formulated the two-hit hypothesis in 1971. Based on a statistical analysis of 48 cases of retinoblastoma, Knudson concluded that, in that case, cancer was triggered by only two mutational events in a single gene. The differences observed between the hereditary and the sporadic/non-hereditary forms of this neoplasm were explained by the presence of one mutation (or *hit*) in the germline DNA followed by a second somatic *hit* in the case of hereditary forms, while two mutations in somatic cells were present for non-hereditary cases (**Figure 13**). The earlier onset of the hereditary forms was also explained by this fact, since only one mutational event was needed for the development of the disease (Knudson, 1971).

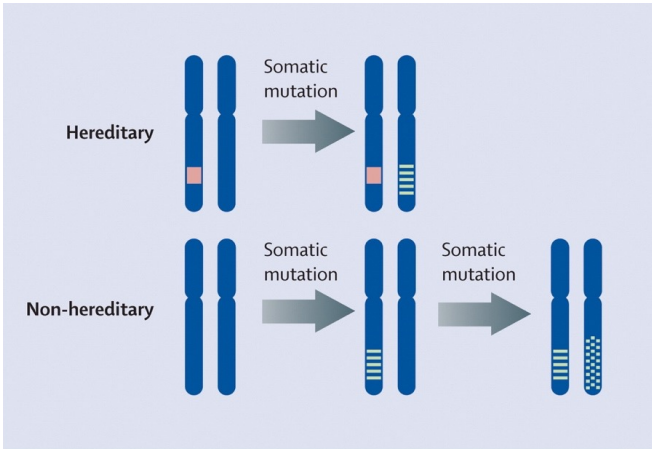


Figure 13. Knudson's two-hit hypothesis. Differences in onset observed between hereditary and non-hereditary cancers are due to a predisposition in form of an inherited germline mutation in the case of hereditary cancer cases (Jozwiak et al., 2008).

This hypothesis was subsequently applied to other cancers and rapidly supported by the scientific community, especially after the setting of the TSG concept. This was the type of genes involved in Knudson's model, since the inactivation of their two alleles was needed for the development of the oncogenic phenotype (Hino & Kobayashi, 2017). Nowadays, we know that any of the first or second hits, leading to the loss of function of a given gene, could take the form of different types of genetic alterations, including SNVs, indels, anomalous methylations or CNVs (mainly second hit deletions leading to losses of heterozygosity (LOHs) of the wild type allele of heterozygous germline alterations).

3.2 Tumor mutational burden

All cancers are characterized by multiple somatic mutations. These mutations can be classified afterwards into driver or passenger mutations according to their effects on tumor development (Stratton, Campbell, & Futreal, 2009). Driver mutations were predominantly prioritized in most cancer sequencing studies because of the growth advantage that they confer, which causes their positive selection during cancer evolution (Stratton et al., 2009; Stratton, 2011). Passenger mutations had so far not been in the spotlight, essentially because they do not confer a selective advantage. However, passenger mutations are also informative, since the total number of passenger and driver mutations allows to extract information both on the number of mitotic cell divisions that occurred in a cell lineage since the fertilized egg and on the mutation rate at each cell division (Stratton, 2011).

This total number of somatic mutations in a tumor genome is called tumor mutational burden (TMB) and it is highly variable among and within cancer types. It ranges from 0.001 mutations per megabase in certain childhood cancers to more than 1,000 according to latest studies (**Figure 14**). Most mutated cancers are those commonly associated with common mutagenic agents such as tobacco smoking (lung cancer) and ultraviolet (UV) light exposure (skin cancer) (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Chalmers et al., 2017). Regarding CRC, it is placed among the top mutated cancers, with some hypermutated cases basically linked to MMR and polymerase epsilon deficiencies, as previously described (Muzny et al., 2012).

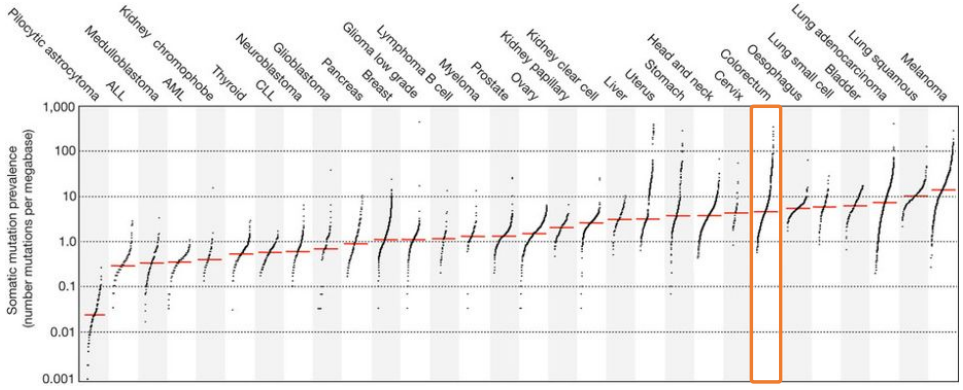


Figure 14. Tumor mutational burden across 30 different cancer types. Number of mutations per megabase identified in different types of human cancer. Every dot represents a cancer sample and red lines represent the median numbers of somatic mutations in each cancer type. Cancer types are ordered on the horizontal axis according to this median tumor mutational burden. Orange box highlights colorectal cancer cases. ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

In recent years, TMB has emerged as a promising biomarker for immunotherapies, due to its association with neoantigen load. It is hypothesized that tumors with high TMB could be more sensitive to activated immune cells since they are more likely to contain neoantigens. These novel cell surface epitopes can be recognized as foreign to the body, thus leading to the increase in T-cell reactivity and the consequent antitumor immune response. In this regard, TMB acts as a convenient biomarker for treatment with immune checkpoint inhibitors, even if not all mutations are likely to induce immunogenic neoantigens (Chalmers et al., 2017; Stenzinger et al., 2019). Immune checkpoint inhibitors, such as inhibitors of cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) and programmed cell death protein 1 (PD-1) receptor, modulate the pathways regulating immune response, increasing antitumor activity by the blockade of immune checkpoints (Pardoll, 2012). Recent clinical studies have associated high TMB with improved patient response and survival rates for these immunotherapy treatments (Yarchoan, Hopkins, & Jaffee, 2017; Chan et al., 2018; Stenzinger et al., 2019).

3.3 Mutational signatures

3.3.1 Formal description

Apart from TMB characterization, passenger mutations found in tumors were also responsible of the emergence of a new exciting field of study in last years. Based on the assumption that the patterns of these passenger mutations are invariable over time, these mutations can be used as a representative picture of the mutational mechanisms that were active during the carcinogenic process (Alexandrov, Nik-Zainal, Wedge, Campbell, & Stratton, 2013). Thus, patterns of driver and passenger mutations reflect the DNA damage and repair processes that cancer cells and their precursors underwent over time (Nik-Zainal et al., 2012).

Each specific mutational process leaves a particular imprint in the genome of a cell, also called *mutational signature* (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). Endogenous cellular mechanisms, such as DNA replication and repair, can generate mutations due to their intrinsic slight infidelity. However, mutations can also arise from exogenous mutagenic exposures. The final record of accumulated DNA damage is determined by the intensity and duration of all active mutational processes (**Figure 15**) (Nik-Zainal et al., 2012).

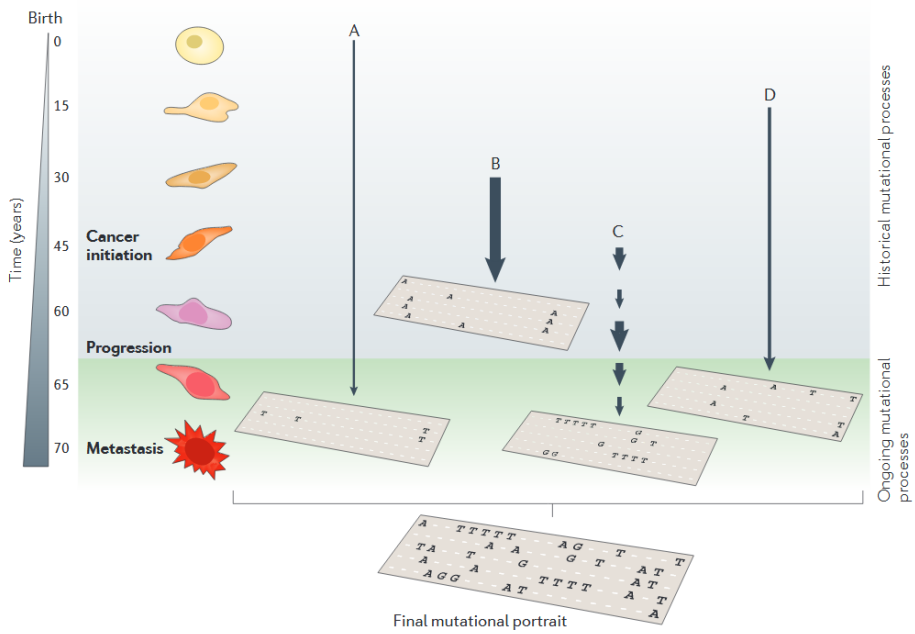


Figure 15. The mutational record of a cancer patient is a mix of different mutational processes. Each mutational source leaves its particular imprint (*mutational signature*) in the genome of a cancer cell, according to the specific duration and intensity of its lifetime exposure (Helleday, Eshtad, & Nik-Zainal, 2014).

DNA damage can emerge in the form of different classes of mutations, such as SNVs, indels or SVs. In order to define a particular structure for the assess of mutational signatures, even if all types of mutations should be considered, at first just the different SNVs according to their composing nucleotides were used, mainly for technical reasons. Thus, the current set of well-established mutational signatures considers six different types of single nucleotide changes, based on the mutated pyrimidine of the Watson-Crick base pair, including four transversions, C>A, C>G, T>A and T>G, and two transitions, C>T and T>C. To better characterize the mutational processes responsible, adjacent nucleotides both in 5′ and 3′ contexts are also considered, accounting for a total of 96 possibilities (6 base substitutions * 4 precedent nucleotides * 4 posterior nucleotides) (**Figure 16**). Thus, each mutational signature is composed by a particular distribution of these 96 potential trinucleotide mutations. This framework also allows that signatures composed by the same classes of substitutions, but in different sequence contexts, can be distinguished (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

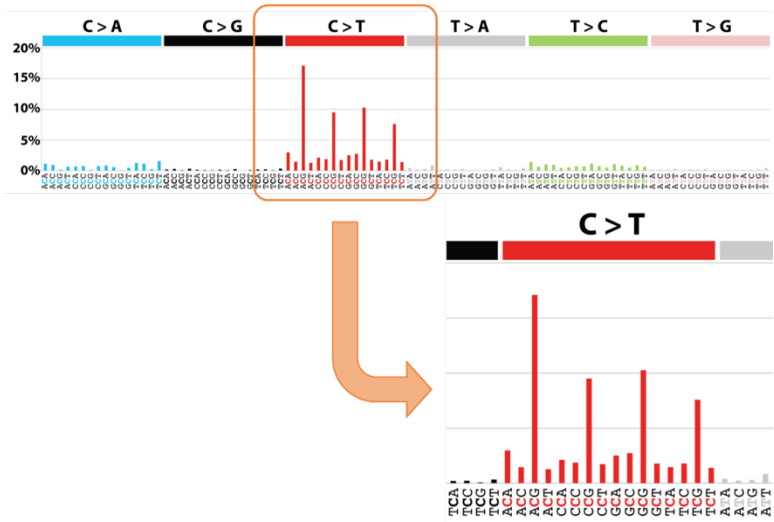


Figure 16. Example of a 96-mutation profile for mutational signatures characterization. Upper panel shows a full profile of the 96 possible single nucleotide variants, considering the 6 possible base substitutions (referred to by the pyrimidine). Lower panel highlights the 16 different possibilities inside every base substitution class, considering both preceding and posterior nucleotides. Mutation profile presented corresponds to signature 1, according to the Catalogue of Somatic Mutations in Cancer database (mutational signatures v2 - March 2015) (Tate et al., 2018; Wellcome Trust Sanger Institute, 2019b).

3.3.2 Computational framework

Even if common features of some of the major mutational patterns could be determined by visual inspection of the profiles, a formal mathematical approach was required in order to improve the quantification of the contribution of each process to the mutational catalog of a specific cancer sample, as well as to allow the identification of subtler alterations leading to specific signatures (Nik-Zainal et al., 2012). A theoretical model of mutational signatures was built as a blind source separation problem and non-negative matrix factorization (NMF) was implemented in order to define an appropriate computational framework (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). This unsupervised machine-learning algorithm is used to extract common features from multidimensional complex data and it had primarily been used for face recognition and text mining (D. D. Lee & Seung, 1999; Berry, Browne, Langville, Pauca, & Plemmons, 2007). However, in recent years, NMF was established as a common approach for different applications in computational biology, including mutational signature analysis (Devarajan, 2008).

3.3.3 Reference mutational signatures

The number of signatures extracted from mutational profiles of published cancer samples evolved as the number of these samples increased, since the number of available genomes and their associated TMB mathematically constrain the number of signatures that can be retrieved by the model (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). In addition, the type of genetic alterations studied in order to decipher new signatures was also increased with the development of the mathematical model and the rise in the number of available samples. A first attempt to extract somatic mutational signatures was performed on 21 breast cancer WGS samples. Five signatures linked to SNVs were deciphered (Nik-Zainal et al., 2012), that subsequently were reduced to four after a further refinement of the approach and its computational implementation (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Thereafter, in a seminal study, Alexandrov and collaborators analyzed 4,938,362 somatic mutations from 7,042 samples of 30 different cancer types, which led to the extraction of 21 different mutational signatures. Most common factors linked to mutational processes were associated to a specific signature, as in the case of aging, tobacco smoking, UV light exposure or defective DNA MMR, whereas approximately half of the identified signatures remained with an unknown etiology. The patterns of contributions to individual cancer samples varied markedly between signatures even within the same cancer type. Some of them contributed a relatively similar number of mutations to most cancers, whereas others gave rise to an important number of mutations but just in specific samples and cancer sites (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

During last years, the reference set of mutational signatures was composed by a total of 30 mutational signatures, considering only SNVs and derived from the analysis of 12,000 samples from 40 different cancer types (**Figure 17**) (Alexandrov et al., 2015; Tate et al., 2018). This gold standard set was used in most publications considering mutational signature analysis (Grolleman, Díaz-Gay, Franch-Expósito, Castellví-Bel, & de Voer, 2019) and it can still be found as part of the Catalogue of Somatic Mutations in Cancer (COSMIC) database (v2 – March 2015) (Wellcome Trust Sanger Institute, 2019a).

For the current set of reference mutational signatures, 49 SNV signatures (also known as single base substitution (SBS) signatures) were extracted from a total of more than 23,000 samples of most cancer types, including over 4,500 WGS cancer samples (Alexandrov et al., 2019). They are indexed in the newest version of COSMIC (v3 – May 2019), along with 17 indel signatures and 11 additional signatures linked to doublet base substitutions (DBSs) (**Figure 18**) (Tate et al., 2018; Wellcome Trust Sanger Institute, 2019b). In this latter case, new specific profiles were developed in order to subclassify these mutational classes (consisting of 83 and 78 mutational subtypes for indels and DBSs respectively) (Alexandrov et al., 2019).

Figure 17. Reference set of mutational signatures from March 2015 until May 2019. It is composed by 30 signatures for single nucleotide variants (subclassified in a profile of 96 subtypes). It can be retrieved in the Catalogue of Somatic Mutations in Cancer database (mutational signatures v2 - March 2015) (Wellcome Trust Sanger Institute, 2019a).

Regarding additional mutational classes, CNV and SV signatures will probably be included in future updates of the framework. Some attempts have been made in this regard for SVs in some recent studies, with DNA repair deficiencies dominating the spectrum of mutational processes responsible for these signatures (Nik-Zainal et al., 2016; Macintyre et al., 2018). However, as for now, signature analysis for CNVs and SVs remains a computational challenge and, in order to be broadly implemented, consensus for their computation must be achieved. Additionally, expanded penta- and heptanucleotide contexts can also be used for point mutations, allowing the extraction of new specific mutational signatures. In this case, not only the adjacent nucleotides are used for characterizing the mutation, but also the two or three possible preceding and posterior nucleotides (Alexandrov et al., 2019).

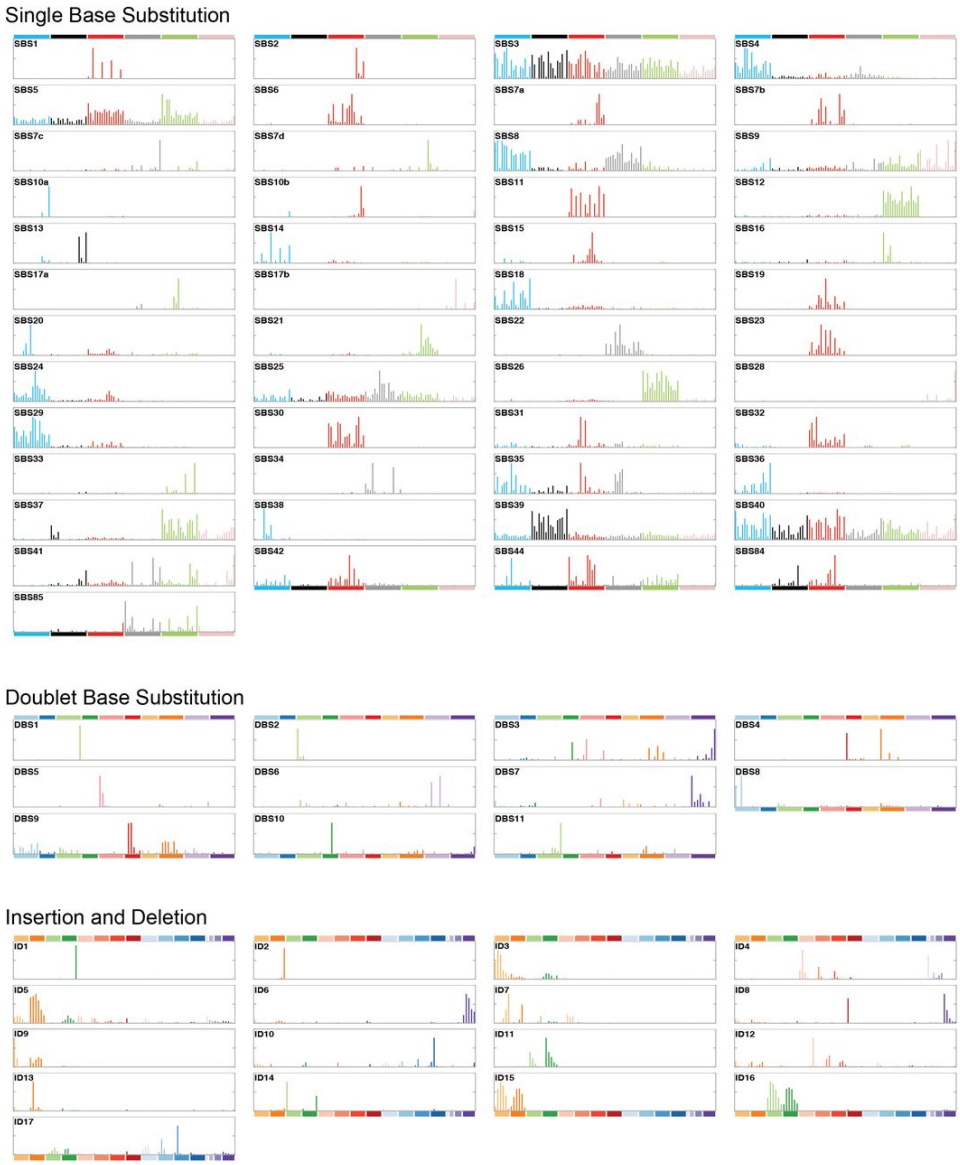


Figure 18. Current set of reference mutational signatures for single base substitutions, doublet base substitutions and small insertions and deletions. It is composed by 47 signatures for single base substitutions (subclassified in a profile of 96 subtypes), 11 doublet base substitutions signatures (78-subtypes profile) and 17 indel signatures (83-subtypes profile) (Alexandrov et al., 2019; Wellcome Trust Sanger Institute, 2019b).

3.3.4 Mutational signatures in colorectal cancer

Different tissues present different cell replacement turnover ratios (Alexandrov et al., 2015). Along with a distinct influence of environmental exposures and other mutational sources, this fact results in a particular distribution of mutational signatures prevalence among tissues. Additionally, it is also plausible that variations of the same signatures could exist between tissues, not already identified due to the mathematical model used so far and the lack of statistical power (Alexandrov et al., 2019). Turnover ratio differences are reflected by those signatures accounting for mutations since the fertilized egg in a steady manner. This is the case of the so-called *clock-like* mutational signatures (SNV signatures SBS1 and SBS5), which therefore reflect the influence of the aging process in the carcinogenesis (Alexandrov et al., 2015). Specifically, signature SBS1 is attributed to spontaneous deamination of 5-methylcytosine at NpCpG trinucleotides, leading to T-G mismatches not repaired before DNA replication and resulting in C>T transitions (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). Even if this signature is present in the mutational catalogue of almost every cancer genome, it is responsible for a different number of mutations depending on the cancer type. Since those differences cannot be attributable to variations in CpG methylation across cell types (Horvath, 2013), they would be reflecting differences in mitotic rates among tissues (Alexandrov et al., 2015). The second mutational signature related with aging, SBS5, is a relatively *flat* signature, i.e., with a uniform distribution of the mutations among the standard 96 possibilities of the mutational profiles. It also presents transcriptional strand bias for T>C transitions in ApTpN context. However, the biological insight underlying for the mutations linked to signature SBS5 is not well understood yet (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Alexandrov et al., 2015). In addition, in the latest set of reference mutational signatures, a novel signature (SBS40), highly similar to SBS5, was also found associated with the aging process (Alexandrov et al., 2019).

Regarding CRC, according to the information displayed in COSMIC and different recently published studies, the interplay among a number of mutational signatures have hitherto been validated (Tate et al., 2018; Wellcome Trust Sanger Institute, 2019b). As previously mentioned, signatures linked to the process of aging (SBS1, SBS5 and recently SBS40) are present in almost every cancer type, therefore including CRC. However, the number of mutations derived from these *clock-like* mutational process is usually low, especially in comparison with those signatures related with two well-known molecular defects present in CRC: DNA MMR and polymerase epsilon proofreading deficiencies. Malfunctioning of both DNA repair pathways leads to the presence of additional mutational signatures implicated in colorectal carcinogenesis. In this regard, signatures SBS10a and SBS10b are strongly associated with mutations in the exonuclease domain of *POLE* gene, whereas in the case of MMR, up to 7 mutational

signatures are linked to this genetic defect (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44) (Alexandrov et al., 2019). In fact, SBS14 is actually associated with concurrent somatic *POLE* mutation and MMR deficiency, whereas in the case of SBS20 it is the mutation of *POLD1* which takes place at the same time as the defect in the MMR pathway (Haradhvala et al., 2018). Signature SBS20 has also been observed in a *MLH1* knockout organoid and is therefore associated with MMR deficiency by *MLH1* inactivation (Drost et al., 2017). All MMR-related mutational signatures were found active in CRC, except from SBS14, according to latest studies (Alexandrov et al., 2019). In the case of SBS20 and SBS26, they were often found in combination with signature SBS6 in the same samples (Nagahashi et al., 2016).

Recently, the profile of mutational signatures affecting colorectal neoplasia has been updated with another main DNA repair pathway. This is the case of BER, and its two main genes closely related with CRC predisposition: *MUTYH* and *NTHL1*. *MUTYH* is a DNA glycosylase gene playing a fundamental role in the repair of oxidative DNA damage. Biallelic germline inactivation of *MUTYH* leads to the well-known predisposition syndrome MAP (Al-Tassan et al., 2002; Weren et al., 2018). Two different studies reported a link between the specific pattern of somatic mutations of MAP patients and a specific mutational signature dominated by C>A mutations, mainly in a CpA context (Pilati et al., 2017; Viel et al., 2017). However, they disagreed in the signature associated, signature SBS18 (Pilati et al., 2017) and signature SBS36 (Viel et al., 2017) respectively, although they closely resemble each other (Pearson correlation coefficient of 0.77) (Viel et al., 2017). In the new framework validated for all cancer types, signature SBS36 was established as the signature associated with *MUTYH* deficiency, whereas signature SBS18 was linked to DNA damage caused by reactive oxygen species (ROS) (closely related with the activity of BER pathway) (Alexandrov et al., 2019). In fact, signature SBS18 was also found in *in vitro* cell models, possibly as the result of induced oxoG damage (Blokzijl et al., 2016; Drost et al., 2017). On the other hand, *NTHL1* is also a DNA glycosylase which have recently been linked with germline predisposition to adenomatous polyposis and CRC (Weren, Ligtenberg, et al., 2015). In this case, stem cell organoids modified by the CRISPR-Cas9 (from Clustered Regularly Interspaced Short Palindromic Repeats – CRISPR-associated protein 9) technique, have been used to functionally validate the link between *NTHL1* deficiency and signature SBS30, characterized by C>T mutations at non-CpG sites (Drost et al., 2017; Grolleman, de Voer, et al., 2019). In addition, the mutational pattern of a breast cancer sample harboring a germline pathogenic variant in *NTHL1* and subsequent somatic LOH closely resembled the profile found in *NTHL1* knockout organoids (Nik-Zainal et al., 2016; Drost et al., 2017). Thus, along with a recent study broadening the spectra of cancers developed by biallelic germline *NTHL1* mutations (Grolleman, de Voer, et al., 2019), the link of this gene deficiency and signature SBS30 was also confirmed for additional neoplasias.

Mutational signatures SBS17a and SBS17b (previously known as a unique signature 17 in version 2 of the mutational signatures in COSMIC) were also recently identified as new signatures implicated in colorectal carcinogenesis (Roerink et al., 2018; Alexandrov et al., 2019). Validated etiology for these signatures have not been reported yet, although 8-hydroxydeoxyguanosine triphosphate induced by ROS was suggested to play a role, leading to a mutational profile enriched in A>C transversions (Dulak et al., 2013). More recently, other signatures have been linked to CRC with a less prevalent role, as it is the case of SBS2 and SBS13 (related to the activity of the apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) family of cytidine deaminases), SBS3 (defective homologous recombination DNA repair and mutation of *BRCA1* and *BRCA2* genes), SBS9 (polymerase eta activity), SBS18 (ROS), SBS12, SBS28, SBS37 and SBS41 (unknown etiology) and even a new signature predominantly characterized by T>G, T>A and T>C mutations at NpTpA and NpTpT trinucleotides and identified by Roerink and collaborators in a study of CRC at single-cell level (Nagahashi et al., 2016; Roerink et al., 2018; Alexandrov et al., 2019).

3.3.5 Mutational signatures as a tool to identify genetic predisposition

Mutational signatures, as well as TMB, can be used for the identification of germline genetic defects that have been active during the origin and evolution of a cancer. This was particularly evident for mutational signatures with a known etiology, and in particular for those associated to germline cancer predisposition syndromes derived from DNA repair deficiencies (**Figure 19**) (J. Ma, Setton, Lee, Riaz, & Powell, 2018; Van Hoeck, Tjoonk, van Boxtel, & Cuppen, 2019). Several tumors arising from pathogenic germline mutations in MMR genes, *BRCA1/2*, *POLE*, *POLD1*, *MUTYH* or *NTHL1*, among others, were identified to be mainly characterized by the associated signatures on their somatic mutational profiling analysis (Campbell et al., 2017; Davies et al., 2017; Viel et al., 2017; Ahadova et al., 2018; Castellsagué et al., 2019; Grolleman, de Voer, et al., 2019).

In the case of polymerase proofreading and MMR germline deficiencies it is also noteworthy the link with high TMB, leading to hypermutated tumors, that has previously been described in colorectal, endometrial and other cancers (Muzny et al., 2012; Kandoth et al., 2013). This hypermutation was defined for a TMB over 10 mutations per megabase. Even if this phenotype could be linked to exogenous mutational processes, in those cases with greater number of mutations (called ultrahypermutated tumors; > 100 mutations per megabase) a mutation in either *POLE/POLD1* or in any of the MMR genes was found. However, in this regard mutational signature analysis could give an extra light, since the analysis of the mutation profile allows to decipher the specific deficiency leading to the hypermutated phenotype (Campbell et al., 2017).

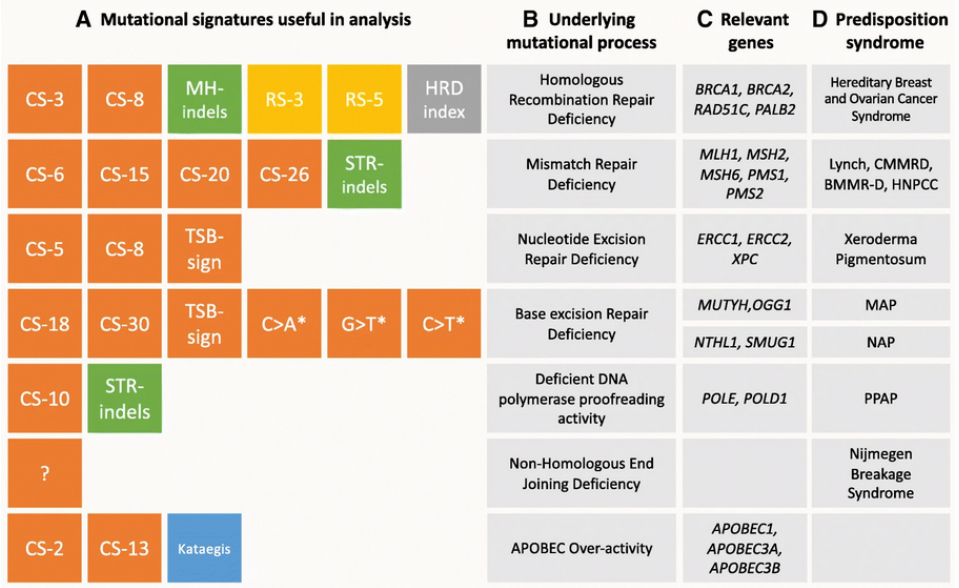


Figure 19. Cancer predisposition syndromes association with mutational processes and associated mutational signatures. Current interrelationship among different DNA repair deficiencies and their related genes, predisposition syndromes and mutational signatures. Signatures derived from different variant classes are considered, namely single base substitutions (orange), indels (green), copy number variants (using the HRD index, described in Watkins, Irshad, Grigoriadis, & Tutt, 2014) (gray) and rearrangements/structural variants (yellow), as well as other molecular markers, such as transcriptional strand bias (orange) or localized hypermutation (known as *kataegis*) (blue). *, defects in base excision repair were associated with these particular single base substitutions; APOBEC, apolipoprotein B mRNA editing catalytic polypeptide-like; BMMR-D, biallelic mismatch repair deficiency; CMMRD, constitutional mismatch repair deficiency; CS, COSMIC signature; HNPCC, hereditary non-polyposis colorectal cancer; HRD, homologous recombination deficiency; indel, small insertion or deletion; MAP, *MUYTH*-associated polyposis; NAP, *NTHL1*-associated polyposis; MH, microhomologies; PPAP, polymerase proofreading-associated polyposis; RS, rearrangement signature; sign, signatures; STR, short tandem repeats; TSB, transcriptional strand bias. Adapted from Van Hoeck et al., 2019.

Apart from the detection of known pathogenic variants, the identification of a mutational signature linked to a DNA repair defect can help in the determination of the potential pathogenicity of a new variant detected in a well-known predisposition gene. This was the case for *POLE*, with a novel pathogenic missense mutation in the exonuclease domain (c.833C>A; p.Thr278Lys). In this case, the tumor harboring the mutation showed both hypermutation and predominance of signature 10 (according to COSMIC mutational signatures v2), linked with *POLE* deficiency, thus suggesting the pathogenic potential of the newly identified variant. This behavior was subsequently functionally validated in yeast (Castellsagué et al., 2019).

In addition, new mutational signatures and those not previously associated with a specific cancer type can also be linked to a particular predisposition syndrome. This could be indicating that new biological mechanisms would be implicated in the carcinogenic process, and therefore new signaling pathways and new genes. Thus, mutational signature analysis in tumors can provide a great impact in the genetic diagnosis of patients, as a novel tool to corroborate that cancers are caused by a genetic predisposition. In the case of CRC, a recent study showed a success in this signature-predisposition syndrome correlation (Grolleman, de Voer, et al., 2019), considering the link of *NTHL1* gene deficiency and signature SBS30 that was previously functionally validated in organoids (Drost et al., 2017). Fourteen tumors from seven tissue types (six extracolonic tissues) were analyzed, finding signature SBS30 as the main contributor to the somatic mutational profile in all but one of the tumors. The only tumor without predominance of signature SBS30 was a urothelial cell cancer (UCC), in which signature SBS2 was the most relevant. This signature is one of the most common found in sporadic UCCs, indicating that this tumor had probably a sporadic origin. On the other hand, the prominent role of signature SBS30 in the rest of tumors showed that *NTHL1* deficiency was affecting a broader spectra of tumor types, apart from CRC and polyposis (Grolleman, de Voer, et al., 2019). Additionally, for CRC a putative future example could be the case of signature SBS3, recently found in CRC (Alexandrov et al., 2019). This signature is associated to defective homologous recombination DNA repair, *BRCA1/2* deficiency and also to mutations in *PALB2* according to previous studies (Polak et al., 2017). Interestingly, this gene was recently proposed as a new CRC predisposition gene (AlDubayan et al., 2018).

3.3.6 Software available to perform mutational signature analysis

In recent years, different software packages have been released in order to practically implement mutational signatures analyses (Baez-Ortega & Gori, 2019; Grolleman, Díaz-Gay, et al., 2019; Hanane, Gianluca, & Vittorio, 2019). It is important to distinguish between those tools that allow the deciphering of *de novo* mutational signatures corresponding to a specific set of samples from those that reconstruct the known mutational profiles of the samples using a given set of signatures. In this latter strategy, also known as signature *refitting*, reconstruction is based on a collection of established mutational signatures, which is commonly the one present in COSMIC database. During last years, COSMIC mutational signatures v2 (March 2015) (Wellcome Trust Sanger Institute, 2019a) have been used as the reference dataset in most softwares, even though a transition to new v3 framework (May 2019) (Wellcome Trust Sanger Institute, 2019b), comprising a larger number of signatures and variant classes considered, is expected in coming years. Hitherto, *de novo* signature extraction was used by most international studies of different cancer types, since it was essential to decipher

which mutational signatures were contributing to the mutational catalogue of every specific tissue type. However, signature refitting could have great potential considering future clinical applications of this methodology. In this regard, limiting factors of *de novo* approaches, such as computational resources and processing time, can be overcome by the possibility to perform a more clinically oriented sample by sample refitting analysis according to a set of consensus signatures (Rosenthal, McGranahan, Herrero, Taylor, & Swanton, 2016; Baez-Ortega & Gori, 2019). Thus, mutational signature analysis could have a big impact in cancer diagnosis, prognosis and treatment, as it has been initially shown for example in the case of *BRCA1/2* and *NTHL1* deficiencies (Davies et al., 2017; Grolleman, de Voer, et al., 2019)

Regarding the extraction of *de novo* signatures, the original implementation of mutational signatures NMF-based computational framework was developed using MATLAB (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). It is currently known as SigProfiler and is available at http://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler. SigProfiler was used in many of the initial studies considering mutational signatures (Grolleman, Díaz-Gay, et al., 2019). However, the use of a proprietary programming language limited its widespread use. To overcome this situation, translation of the original framework to open access Python language has been done in last years (Alexandrov et al., 2019). On the other hand, alternative softwares were built using different open source programming languages and platforms. The R programming language, reference for bioinformatic community, was used by SomaticSignatures (Gehring, Fischer, Lawrence, & Huber, 2015), MutationalPatterns (Blokzijl, Janssen, van Boxtel, & Cuppen, 2018), mutSignatures (Fantini et al., 2018), Palimpsest (Shinde et al., 2018), Maftools (Mayakonda, Lin, Assenov, Plass, & Koeffler, 2018), Helmsman (Carlson, Li, & Zöllner, 2018) and the soon to be published CANCERSIGN (Bayati et al., 2019), all of them mimicking the NMF model to extract mutational signatures, although with different additional specific features. Mutational signatures computational framework was also transposed to the well-known bioinformatic platform Galaxy, via the MutSpec toolbox (Ardin et al., 2016). Alternatively, some other computational approaches were developed using different strategies outside the NMF paradigm or adding some additional concepts to the algorithm. This was the case of probabilistic models of both EMu (Fischer, Illingworth, Campbell, & Mustonen, 2013) and pmsignature (probabilistic mutation signature) (Shiraishi, Tremmel, Miyano, & Stephens, 2015), the empirical Bayesian approach of signeR (Rosales, Drummond, Valieris, Dias-Neto, & da Silva, 2017), a multi-modal correlated topic model (Funnell et al., 2019) and also two future releases based on Bayesian inference on probabilistic signature models (sigfit) (Gori & Baez-Ortega, 2018) and on NMF with a Lasso-penalized cost function (SparseSignatures) (Ramazzotti, Lal, Liu, Tibshirani, & Sidow, 2019), respectively.

Particularly, a Bayesian variant of NMF was recently implemented to better estimate the number of underlying mutational processes of a particular set of samples, and hence the number of signatures implicated. The so-called SignatureAnalyzer tool was built in R language and successfully applied in different studies (V. Y. F. Tan & Fevotte, 2013; Kasar et al., 2015; Kim et al., 2016; Haradhvala et al., 2018). Along with the original SigProfiler implementation, SignatureAnalyzer has been used as the computational framework to deconvolute the new consensus repertoire of mutational signatures across all cancer types (Alexandrov et al., 2019).

Conversely, a different computational challenge arises with respect to the reconstruction of mutational profiles using a set of consensus mutational signatures. The first available tool of this class was deconstructSigs, using iterative multiple linear regression for extracting the contributions of known mutational signatures (Rosenthal et al., 2016) and being widely used since its publication (Bruna et al., 2016; Goh et al., 2016; Hao et al., 2016; Kanu et al., 2016; Nagahashi et al., 2016). Additionally, signature refitting was allowed by some of the softwares also performing *de novo* deciphering. This kind of comprehensive approach was implemented in MutationalPatterns (Blokzijl et al., 2018) , where a non-negative least squares (NNLS) approximation was used to extract the contributions of reference signatures at sample resolution in some recent reports (Blokzijl et al., 2016; Drost et al., 2017). In addition, MutationalPatterns achieved a substantial enhancement in computation time with respect to deconstructSigs, since its analysis runtime is approximately 400 times faster (Blokzijl et al., 2018). Apart from these two tools, other softwares were developed for signature refitting, including YAPSA, an R package using linear combination decomposition (Huebschmann, Gu, & Schlesner, 2019), MutationalCone, a future release based on cone projection (Hanane et al., 2019), and quadratic programming-based packages QPsig (A. G. Lynch, 2016), decompTumor2Sig (Krüger & Piro, 2019) and SignatureEstimation, which used both quadratic programming and simulated annealing optimization techniques (X. Huang, Wojtowicz, & Przytycka, 2018). Even the original SigProfiler computational framework presented a specific implementation, known as SigProfilerSingleSample, which has also been recently transposed to Python (Alexandrov et al., 2015, 2019).

Regardless of the vast amount of options available to perform mutational signature analysis, for an important part of the scientific community it remains inaccessible. Along with the need for substantial computing capacity, especially in the case of the analysis of large cohorts, software packages developed are predominantly useful for bioinformatic experts and should be adapted to existing somatic analysis pipelines.

HYPOTHESIS | OBJECTIVES

Hypothesis

Colorectal cancer (CRC) is a complex disease, thus with an etiology mixing both genetic and environmental factors. Genetic predisposition encompasses up to 35% of CRCs according to twin and family studies, whereas the well-known predisposition syndromes linked to specific germline defects only explain 2-8% of cases. Therefore, a missing heritability is present for this neoplasm. Next generation sequencing is the most suitable tool to address the identification of new genes implicated in disease predisposition, as it has been proved in recent studies involving genes such as *POLD1*, *POLE* or *NTHL1*. However, this technology identifies a large number of genetic variants in every patient, thus generating the need of a prioritization strategy. In this regard, not only germline but also somatic genetic alterations can play a fundamental role in providing new insights on CRC hereditary predisposition, in accordance with the classic Knudson's two-hit hypothesis. Accordingly, somatic mutational profile analysis has been recently used for the identification of new CRC predisposition genes, as well as a promising biomarker for diagnosis, prognosis and treatment of this neoplasia. Even though some bioinformatic packages have been developed to address this analysis, it remains inaccessible for a substantial proportion of the scientific community.

Objectives

General objective

The main purpose of this doctoral thesis is to identify novel candidate genes that could be implicated in germline predisposition to familial CRC. A combined germline-tumor whole exome sequencing (WES) data analysis and a bioinformatic application for somatic mutational profiling will be developed to be used as prioritization approaches.

Specific objectives

1. Development of a computational application to address somatic mutational profile analysis using the Shiny framework of R programming language, in a user-friendly and freely available web environment suitable for non-specialized researchers. Characterization of tumor mutational burden and mutational signatures refitting according to COSMIC reference signatures v2 will be available, as well as sample classification by clustering and principal component analysis.

2. Integrative analysis based on Knudson's two-hit hypothesis of WES data from germline and tumor DNA of a cohort of 18 familial CRC patients, in order to identify new potential tumor suppressor genes. Different classes of genetic alterations, such as single nucleotide variants, small insertions and deletions, copy number variants and losses of heterozygosity, will be considered. Candidate genes will be selected when both germline and tumor DNA are affected by one of these genetic alterations.

3. Somatic mutational profiling of the mentioned cohort of familial CRC using the bioinformatic tool developed previously, considering tumor mutational burden and mutational signatures characterization.

RESULTS

Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples.

Marcos Díaz Gay, Maria Vila Casadesús, Sebastià Franch Expósito, Eva Hernández Illán, Juan José Lozano and Sergi Castellví Bel.

BMC Bioinformatics 2018;19(1):224.

https://doi.org/10.1186/s12859-018-2234-y

Díaz-Gay et al. BMC Bioinformatics (2018) 19:224
https://doi.org/10.1186/s12859-018-2234-y

BMC Bioinformatics

SOFTWARE Open Access

Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples

Marcos Díaz-Gay[1†], Maria Vila-Casadesús[2,3†], Sebastià Franch-Expósito[1†], Eva Hernández-Illán[1], Juan José Lozano[2] and Sergi Castellví-Bel[1*]

Abstract

Background: Mutational signatures have been proved as a valuable pattern in somatic genomics, mainly regarding cancer, with a potential application as a biomarker in clinical practice. Up to now, several bioinformatic packages to address this topic have been developed in different languages/platforms. MutationalPatterns has arisen as the most efficient tool for the comparison with the signatures currently reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. However, the analysis of mutational signatures is nowadays restricted to a small community of bioinformatic experts.

Results: In this work we present Mutational Signatures in Cancer (MuSiCa), a new web tool based on MutationalPatterns and built using the Shiny framework in R language. By means of a simple interface suited to non-specialized researchers, it provides a comprehensive analysis of the somatic mutational status of the supplied cancer samples. It permits characterizing the profile and burden of mutations, as well as quantifying COSMIC-reported mutational signatures. It also allows classifying samples according to the above signature contributions.

Conclusions: MuSiCa is a helpful web application to characterize mutational signatures in cancer samples. It is accessible online at http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html and source code is freely available at https://github.com/marcos-diazg/musica.

Keywords: Mutational signatures, COSMIC database, Single nucleotide variants, Cancer genomics, Web tool, Shiny, R language

Background

Mutational processes in somatic cells are mainly led by endogenous or exogenous mutagenic agents, as well as errors in DNA replication or repair machineries. Any type of agent or defect is responsible for a specific footprint in the form of a different burden and pattern of mutations. Some of them are historically well-known, as in the case of ultraviolet light exposure and its association with C > T and CC > TT substitutions caused by pyrimidine dimers [1].

In recent years, a new methodology has arisen on this field. Mutational signatures framework enables the association of patterns of mutations with cellular processes and external agents causing them [2]. Since all cancers are caused by somatic mutations, this methodology has the potential to provide insight into their underlying biological processes and become a biomarker in clinical practice [3]. It is based on a computational implementation of non-negative matrix factorization (NMF) considering more than 10,000 cancer samples [4, 5]. Using the information of somatic single nucleotide variants (SNVs), a series of mutational profiles are extracted. These profiles take into account not only substituted nucleotides (all replacements are referred to by the pyrimidine of the

* Correspondence: sbel@clinic.cat
†Marcos Díaz-Gay, Maria Vila-Casadesús and Sebastià Franch-Expósito contributed equally to this work.
[1]Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD, University of Barcelona, Barcelona, Spain
Full list of author information is available at the end of the article

mutated Watson-Crick base pair) but also the 5′ and 3′ adjacent bases. A total of 96 possibilities are evaluated, allowing to detect processes responsible for the same substitutions but in different contexts. According to the current information of the Catalogue of Somatic Mutations in Cancer (COSMIC) database [6], thirty mutational signatures have already been identified across 40 different types of human cancer. This methodology has the potential to reconstruct the mutational spectrum of any cancer sample with sufficient accuracy. This reconstruction is based on the combination of the different signatures contributions. Thus, it constitutes the imprint on the genome of specific mutagenic agents or genetic defects, each represented by a specific signature.

Several bioinformatic approaches have been developed to address mutational signature analysis using different platforms and programming languages. Including some commonalities such as the 96-mutation profile plotting (6 different nucleotide substitutions * 16 different 3-mer contexts), different packages have been recently developed for de novo signature extraction and contribution of known signatures. This is extremely important regarding the possibility of using this methodology in clinical practice. In this context, it would be convenient to perform the analysis at sample resolution, and this is only achievable by comparison with a set of established signatures. MutationalPatterns is an R/Bioconductor package that covers the whole spectrum of functionalities required for mutational signatures framework implementation [7]. It allows the extraction of de novo signatures using the original NMF algorithm, like former R packages pmsignature [8] and Somatic Signatures [9], and Galaxy tool MutSpec [10]. In addition, it also permits the quantification of COSMIC-reported signatures by finding their optimal linear combination. This process is performed approximately 400 times faster than deconstructSigs [11], the only package also covering this functionality [7]. MutationalPatterns has been proved useful in recent studies, both in the identification of somatic mutational profiles [12] and in the characterization of known mutational signatures in human stem cells [13].

However, analysis of somatic mutational signatures remains currently inaccessible for a substantial proportion of the scientific community. Developed software is only useful for bioinformatic experts that should adapt it to their somatic analysis pipelines. Computational resources are also a big challenge, especially when the number of samples to handle is considerably high. In this regard, we have developed a web application to overcome these challenges. Mutational Signatures in Cancer (MuSiCa) allows an easy and quick analysis of mutational signatures in cancer samples, based on a user-friendly web environment adapted to the whole research community. It is mainly built on top of the MutationalPatterns package, so benefiting from its functionalities but also adding a graphical interface designed for non-specialized researchers. MuSiCa also presents some extra features specially designed for cancer samples characterization. Its main aim is to quantify known mutational signatures contribution at sample level, therefore facilitating the identification of the underlying mutational processes. The application also permits to perform an analysis for a complete cohort of cancer patients.

Implementation

MuSiCa was developed using the Shiny framework, which enables the straightforward building of interactive web applications directly from R code [14]. It integrates different publicly available R packages in order to generate a convenient interface and computational efficiency to fluently handle somatic mutation data. Regarding mutational signature framework, MuSiCa uses the data available in COSMIC. Hitherto, the 30 signatures that have been validated and reported in this database have been considered, with the prospect of a future update which would also be transferred to the application. MuSiCa can be easily run online at http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html. Source R code is freely available to download at https://github.com/marcos--diazg/musica, where the required dependencies to install the application are indicated.

A typical workflow of MuSiCa application is presented in Fig. 1. It starts with the uploading of the files containing the somatic SNVs of the samples to analyze. Samples may be derived from international studies as ICGC/TCGA or directly provided by the users. The minimum information required is the chromosome and genomic position according to the human reference genome (UCSC GRCh38/hg38, GRCh37/hg19 and 1000genomes hs37d5 builds are supported), as well as the reference and alternative alleles for every mutation. Different file formats are permitted including the default for this kind of data, the Variant Call Format (VCF). Tab-Separated Values (TSV), Excel and Mutation Annotation Format (MAF) are also allowed. MAF format is commonly used for packing multi-sample data from the Genomics Data Commons projects. Multiple file uploading is allowed in the case of VCF, TSV and Excel formats, each containing the somatic mutations of one sample at a time. For MAF format, only one multi-sample file is allowed. A help modal is present in the MuSiCa website to clarify input format options to the users. The human reference genome build and the type of genomic study performed also need to be provided in order to correctly calculate the prevalence of somatic mutations (i.e. the number of mutations per megabase).

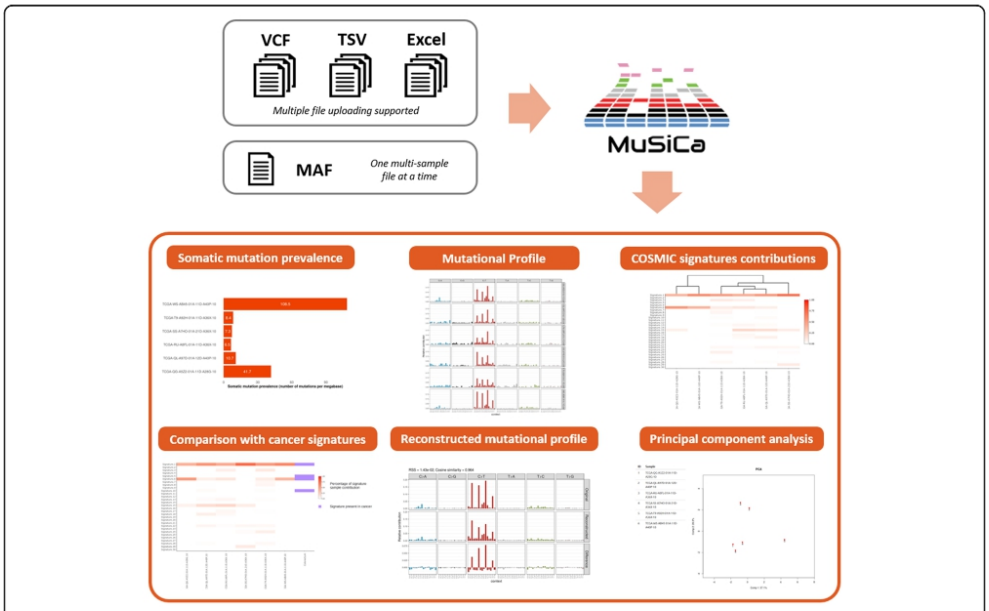


Fig. 1 Analysis workflow of MuSiCa. Input options supported by the web application and output elements available as plots and data tables ready to download by the users

Output elements are displayed in six different tabs. They are presented in the form of publication-ready figures and tables that can be directly downloaded by the users in different formats. Firstly, mutation prevalence and profiling are presented for somatic mutation characterization. Regarding profiles, all possible SNVs considering the substituted base and the 5′ and 3′ adjacent nucleotides are depicted. Regarding the mutational signatures pattern, it is possible to visualize the contribution of COSMIC-reported signatures, as well as those associated with the distinct cancer types present in this database. The application also permits clustering samples and signatures according to the contributions using a distance measure based on Pearson correlation (1 – correlation value), as well as selecting which samples and cancer types are represented. A principal component analysis (PCA) plot is also presented when more than three samples are uploaded. Both clustering and PCA enable the classification of provided samples according to their quantification regarding known mutational signatures.

This process of signatures quantification is based on the least squares method. This method permits to find the optimal linear combination of the 30 signatures that minimize the residual sum of squares (RSS). Therefore, RSS is a measure of the efficiency of the original mutational profile reconstruction. MuSiCa presents an output tab where original and reconstructed profiles are depicted. RSS is also shown, as well as cosine similarity between both profiles. This value presents instead a direct measure of the correspondence between the two depicted profiles in a 0–1 range (identical profiles would have a value of 1). A value above 0.9 is considered as sufficient accuracy.

Results and discussion

To assess the usability of the application, colon cancer SNV data from the NCI Genomic Data Commons was used. Four hundred thirty-three samples of this neoplasia were analyzed. They corresponded to the TCGA-COAD project. Somatic mutation data derived from TCGA projects was freely available in MAF format. As this is one of the supported input formats in MuSiCa, the application permitted to directly analyze this publicly available repository. Different upstream analysis workflows were available, using different somatic variant callers. MuTect2-derived data was selected in this example in accordance with GATK Best Practices [15].

Colorectal cancer is one of leading neoplasms worldwide considering mortality and morbidity. Regarding mutagenic agents, effects of environmental factors such as smoking are well-known. However, defects in key molecular pathways, especially those related with DNA repair, have been established as key factors in this neoplasm. Both malfunctioning of mismatch repair (MMR) genes

and polymerases δ and ε are reported to affect colorectal carcinogenesis [16]. This is particularly important in the case of hypermutated tumors, defined as those having a mutation rate above 12 per 10^6. This malfunctioning could be caused by somatic but also germline genetic alterations. Indeed, Lynch syndrome and Polymerase proofreading-associated polyposis are both hereditary colorectal cancer syndromes related to malfunctioning of previously indicated DNA repair pathways [17].

Results of the analysis of colorectal cancer TCGA samples with MuSiCa are presented in Fig. 2 and Additional file 1. Regarding the quantification of COSMIC signatures, clustering discriminated at least three different subsets of colon cancer samples in this cohort. The group on the left, accounting for more than half of the samples, was mainly characterized by signature 1. This profile has been found in all cancer types and has been correlated with the age of cancer diagnosis. It is produced as an endogenous process derived from spontaneous deamination of 5-methylcytosine. The other two groups presented a higher level of signatures predominantly associated with MMR deficiency (signatures 6, 15 and 20) and defects in polymerase ε (signature 10). This is in agreement with microsatellite-unstable and *POLE*-mutated colon cancers [16]. However, they also showed the impact of age-associated signature 1. Therefore, this is a good example to realize how mutational signatures reconstruction highlighted the impact of the different underlying causes of mutations present in specific cancer samples. This fact could be a key evidence connecting to the carcinogenic process and even germline susceptibility to the neoplasm.

Regarding developed software for mutational signature analysis, some other tools were already available. In reference to bioinformatic packages, some different options were available as previously mentioned. MutationalPatterns has arisen as the most efficient tool enabling the comparison with the currently reported signatures. In recent years, some web applications have also been published in order to improve the accessibility to this methodology to the whole research community. Pmsignature was the first online application ready to apply mutational signatures framework [8]. However, it was intended just to extract new mutational signatures derived from the supplied samples, not allowing the comparison with known signatures. More recent examples include MutaGene, providing a huge computational framework

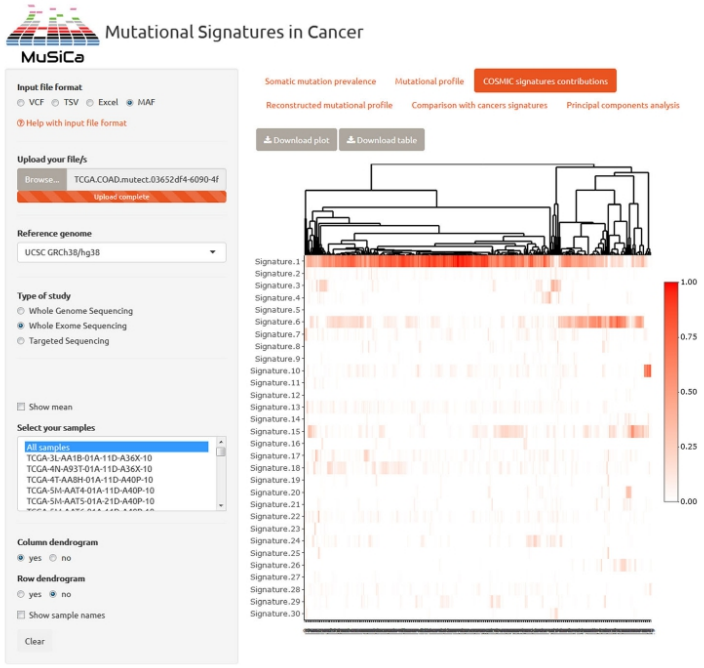


Fig. 2 Capture of MuSiCa web application. Known signatures contribution (reported in COSMIC) of 433 TCGA-COAD colon cancer samples is shown. Input options are on the left-hand side and results appear on the main panel when the analysis is performed

regarding somatic cancer mutations [18]. It includes a large repository regarding mutational signatures, but it is more focused on the analysis of publicly available datasets than samples directly provided by the users. In fact, regarding this last point, it permits analyzing a set of samples but cannot generate analysis reports on a single sample level. mSignatureDB is a recent web implementation that allows for the first time to perform signature analysis on datasets directly uploaded by users [19]. Although it permits to quantify known mutational signatures contributions, it lacks some functionalities regarding sample classification, as clustering or PCA analysis. In addition, quantification process is based on deconstructSigs package, with the mentioned weakness on computational efficiency. To the best of our knowledge, no web application is able to characterize the burden of mutation of different cancer samples, as well as cluster and classify them according to their COSMIC-signatures quantification. Thus, MuSiCa becomes the most comprehensive tool available online for somatic characterization of cancer samples datasets directly provided by users.

Conclusions

Our study shows the potential of the mutational signature framework as a biomarker in cancer and the simplicity and usefulness of our implementation. It is also remarkable that MuSiCa allows the analysis at sample level, which is mandatory regarding future clinical implementation of this methodology. Direct accessibility via web, user-friendly environment and computational performance are key factors of our application.

Availability and requirements

- Project name: MuSiCa
- Project home page: https://github.com/marcos-diazg/musica
- Operating system(s): Platform-independent
- Programming language: R, Shiny
- Other requirements: Internet connectivity
- License: MIT License
- Any restrictions to use by non-academics: No

Additional file

Additional file 1: Figure S1. Somatic mutational prevalence in MuSiCa web app. **Figure S2**. Mutational profile representation in MuSiCa web app. **Figure S3**. Reconstruction of mutational profile in MuSiCa web app. **Figure S4**. Comparison with cancer signatures in MuSiCa web app. **Figure S5**. Principal component analysis in MuSiCa web app. (PDF 3039 kb)

Abbreviations

COSMIC: Catalogue of somatic mutations in cancer; MAF: Mutation annotation format; MMR: Mismatch repair; NMF: Non-negative matrix factorization; PCA: Principal component analysis; RSS: Residual sum of squares; SNV: Single nucleotide variant; TSV: Tab-separated values; VCF: Variant call format

Acknowledgements

We are sincerely grateful to Pau Erola, Hadrián Villar and ATIC-UPC for technical support. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.

Funding

MDG and SFE are supported by contracts from FI 2017 (B00619, AGAUR, Generalitat de Catalunya) and CIBEREHD, respectively. CIBEREHD is funded by the Instituto de Salud Carlos III. This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (14/00173, 17/00878), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), COST Action CA17118, PERIS (SLT002/16/00398, Generalitat de Catalunya), CERCA Programme (Generalitat de Catalunya) and Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya, 2014SGR255, GRPRE 2017SGR21). The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All source code has been made publicly available on Github at: https://github.com/marcos-diazg/musica.

Authors' contributions

MDG, MVC, SFE and SCB conceived the idea. MDG, MVC, SFE, EHI and JJL developed the application. MDG and SCB wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD, University of Barcelona, Barcelona, Spain. [2]Bioinformatics Platform, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Barcelona, Spain. [3]Present Address: Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

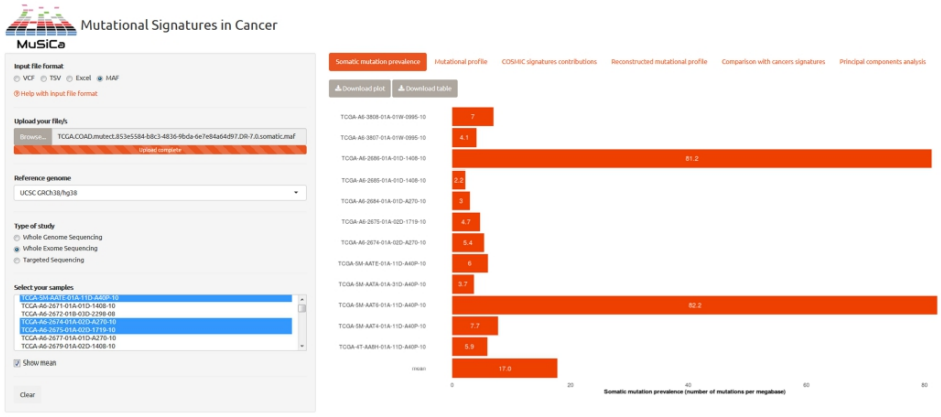
Received: 5 February 2018 Accepted: 4 June 2018
Published online: 14 June 2018

References

1. Miller JH. Mutagenic specificity of ultraviolet light. J Mol Biol. 1985;182:45–65.
2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.
3. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017;23:517–25.
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3:246–59.
5. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.
6. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.

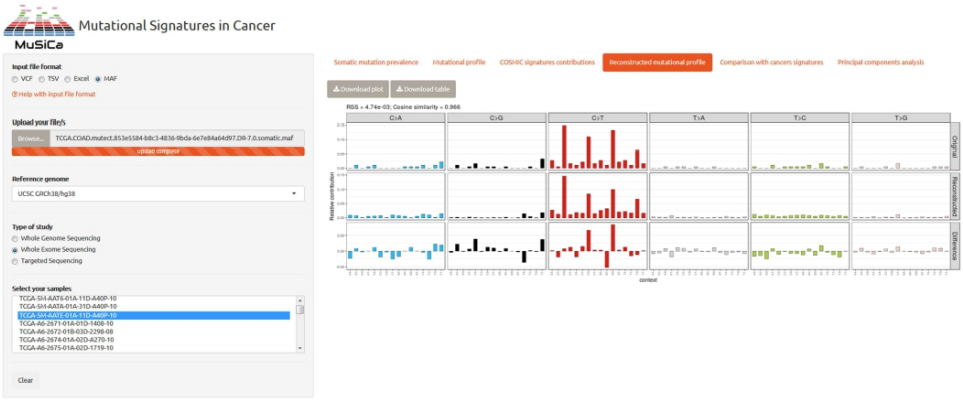
7. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10:33.
8. Shiraishi Y, Tremmel G, Miyano S, Stephens M. A simple model-based approach to inferring and visualizing Cancer mutation signatures. Marchini J, editor. PLOS Genet. 2015;11:e1005657.
9. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. 2015;31:3673–5.
10. Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, et al. MutSpec: a galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinformatics. 2016;17:170.
11. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.
12. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4.
13. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017;358:234–8.
14. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. Shiny: Web Application Framework for R [Internet]. 2017. Available from: https://cran.r-project.org/package=shiny
15. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
16. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
17. Valle L. Recent discoveries in the genetics of familial colorectal Cancer and polyposis. Clin Gastroenterol Hepatol. 2017;15:809–19.
18. Goncearenco A, Rager SL, Li M, Sang Q-X, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Res. 2017;45:W514–22.
19. Huang P-J, Chiu L-Y, Lee C-C, Yeh Y-M, Huang K-Y, Chiu C-H, et al. mSignatureDB: a database for deciphering mutational signatures in human cancers. Nucleic Acids Res. 2018;46:D964–70.

Supplementary Material

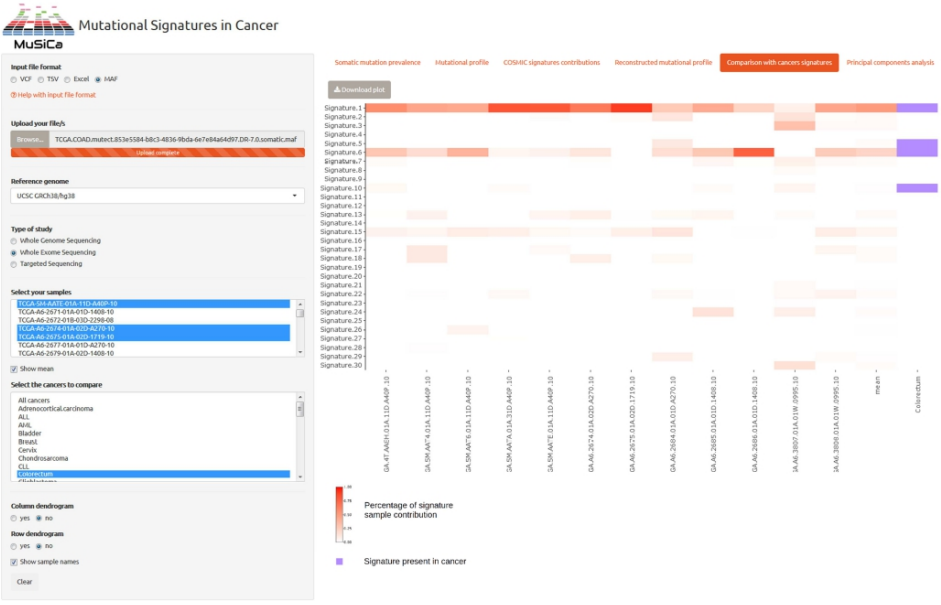


Supplementary Fig. 1. Somatic mutational prevalence in MuSiCa web app. Mutational burden output tab showing a subset of the TCGA-COAD project samples and its mean value.

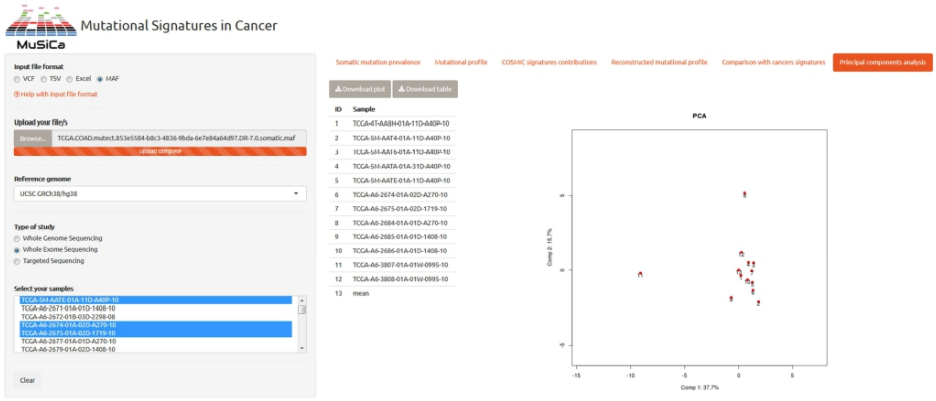
Supplementary Fig. 2. Mutational profile representation in MuSiCa web app. Mutational profile output tab showing a subset of the TCGA-COAD samples.



Supplementary Fig. 3. Reconstruction of mutational profile in MuSiCa web app. Reconstructed mutational profile output tab showing a specific sample of the TCGA-COAD project.



Supplementary Fig. 4. Comparison with cancer signatures in MuSiCa web app. Output tab presenting a comparison of known signatures contributions with mutational signatures reported in different human cancer types in a subset of the TCGA-COAD samples and its mean value.



Supplementary Fig. 5. Principal component analysis in MuSiCa web app. Principal component analysis output tab presenting a classification of a specific subset of the TCGA-COAD samples according to the quantification of known signatures contributions.

Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer.

Marcos Díaz Gay, Sebastià Franch Expósito, Coral Arnau Collell, Solip Park, Fran Supek, Jenifer Muñoz, Laia Bonjoch, Anna Gratacós Mulleras, Paula Alejandra Sánchez Rojas, Clara Esteban Jurado, Teresa Ocaña, Miriam Cuatrecasas, Maria Vila Casadesús, Juan José Lozano, Genís Parra, Steven Laurie, Sergi Beltran, EPICOLON Consortium, Antoni Castells, Luis Bujanda, Joaquín Cubiella, Francesc Balaguer and Sergi Castellví Bel.

Cancers 2019;11(3):362.

https://doi.org/10.3390/cancers11030362

cancers

MDPI

Article

Integrated Analysis of Germline and Tumor DNA Identifies New Candidate Genes Involved in Familial Colorectal Cancer

Marcos Díaz-Gay [1], Sebastià Franch-Expósito [1], Coral Arnau-Collell [1], Solip Park [2], Fran Supek [3], Jenifer Muñoz [1], Laia Bonjoch [1], Anna Gratacós-Mulleras [1], Paula A. Sánchez-Rojas [1], Clara Esteban-Jurado [1], Teresa Ocaña [1], Miriam Cuatrecasas [4], Maria Vila-Casadesús [5], Juan José Lozano [5], Genis Parra [6], Steve Laurie [6], Sergi Beltran [6], EPICOLON Consortium [1,7,8], Antoni Castells [1], Luis Bujanda [7], Joaquín Cubiella [8], Francesc Balaguer [1] and Sergi Castellví-Bel [1,*]

1 Gastroenterology Department, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clínic, 08036 Barcelona, Spain; diaz2@clinic.cat (M.D.-G.); sebasfranch@gmail.com (S.F.-E.); arnau@clinic.cat (C.A.-C.); jenifer.munoz@ciberehd.org (J.M.); bonjoch@clinic.cat (L.B.); annagratacosm@gmail.com (A.G.-M.); paulasanchez10@live.com (P.A.S.-R.); darth.clara@gmail.com (C.E.-J.); mocana@clinic.cat (T.O.); scastellvibel@msn.com (E.C.); castells@clinic.cat (A.C.); fprunes@clinic.cat (F.B.)

2 Systems Biology Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain; imagineyd@gmail.com

3 Institut de Recerca Biomedica (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain; fran.supek@irbbarcelona.org

4 Pathology Department, Hospital Clínic, 08036 Barcelona, Spain; mcuatrec@clinic.cat

5 Bioinformatics Platform, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), 08036 Barcelona, Spain; maria.vila.cs@gmail.com (M.V.-C.); juanjo.lozano@ciberehd.org (J.J.L.)

6 Centre Nacional d'Anàlisi Genòmica-Centre de Regulació Genòmica (CNAG-CRG), Parc Científic de Barcelona, 08028 Barcelona, Spain; genis.parra@cnag.crg.eu (G.P.); steven.laurie@cnag.crg.eu (S.L.); sergi.beltran@cnag.crg.eu (S.B.)

7 Gastroenterology Department, Hospital Donostia-Instituto Biodonostia, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Basque Country University (UPV/EHU), 20014 San Sebastián, Spain; luis.bujandafernandezdepierola@osakidetza.eus

8 Gastroenterology Department, Complexo Hospitalario Universitario de Ourense, Instituto de Investigación Sanitaria Galicia Sur, 32005 Ourense, Spain; joaquin.cubiella.fernandez@sergas.es

* Correspondence: sbel@clinic.cat; Tel.: +34-93227-5400 (ext. 4183)

Received: 25 January 2019; Accepted: 9 March 2019; Published: 13 March 2019

Abstract: Colorectal cancer (CRC) shows aggregation in some families but no alterations in the known hereditary CRC genes. We aimed to identify new candidate genes which are potentially involved in germline predisposition to familial CRC. An integrated analysis of germline and tumor whole-exome sequencing data was performed in 18 unrelated CRC families. Deleterious single nucleotide variants (SNV), short insertions and deletions (indels), copy number variants (CNVs) and loss of heterozygosity (LOH) were assessed as candidates for first germline or second somatic hits. Candidate tumor suppressor genes were selected when alterations were detected in both germline and somatic DNA, fulfilling Knudson's two-hit hypothesis. Somatic mutational profiling and signature analysis were also performed. A series of germline-somatic variant pairs were detected. In all cases, the first hit was presented as a rare SNV/indel, whereas the second hit was either a different SNV (3 genes) or LOH affecting the same gene (141 genes). *BRCA2*, *BLM*, *ERCC2*, *RECQL*, *REV3L* and *RIF1* were among the most promising candidate genes for germline CRC predisposition. The identification of new candidate genes involved in familial CRC could be achieved by our integrated analysis. Further functional studies and replication in additional cohorts are required to confirm the selected candidates.

Keywords: colorectal cancer; whole-exome sequencing; predisposition to disease; germline–tumor analysis; mutational signatures; computational genomics

1. Introduction

Colorectal cancer (CRC) is one of the most common and lethal malignant neoplasms worldwide, accounting for 8% of all cancer-related deaths [1]. Developed countries are the most affected, with almost 55% of diagnosed cases, although with better survival rates, since 52% of deaths occur in less-developed regions [2]. The lifetime risk of developing CRC is between 5–6%, but an incidence rise is expected in the coming years, due to higher life expectancy.

Genetic and environmental factors are involved in CRC predisposition. Environmental contributors include alcohol, tobacco and fat intake, among others [3]. Inherited genetic variation reaches 35% of susceptibility according to twin studies [4]. Predisposition can be classified according to population frequency and associated disease risk into high- and low-penetrant variants. High-penetrant variants are rare and have a large effect on the predisposition to the disease. Regarding CRC, well-defined genes such as *APC*, *MUTYH*, the DNA polymerases *POLE* and *POLD1* and the DNA mismatch repair (MMR) family (*MSH2*, *MLH1*, *MSH6* and *PMS2*) are affected by these mutations, causing well-known hereditary syndromes (familial adenomatous polyposis, *MUTYH*-associated polyposis, polymerase proofreading-associated polyposis and Lynch syndrome, respectively) [5]. However, only 5% of CRC cases are explained by this kind of variation due to its low frequency in the population. Low-penetrance genetic variation, mainly identified in genome-wide association studies, is characterized by a high prevalence in the population and a weak deleterious effect. However, collectively, all identified low-penetrance variants contribute significantly to CRC susceptibility, accounting for 5–10% of the heritability to this disease [6].

Familial CRC can be defined as a heterogeneous condition defined by patients with a family history for this neoplasia without alterations in the known hereditary CRC genes. Its etiology is not completely understood yet. The genes responsible are likely to be fairly uncommon but penetrant enough to explain the autosomal dominant patterns of inheritance reported [6]. Recent studies identified potentially implicated genes, including *NTHL1*, *GREM1* and *RNF43*, as the most remarkable [7].

The two-hit hypothesis for cancer development was formulated by Alfred G. Knudson in 1971 [8]. Genes with a loss of function followed by a rapid acceleration of the oncogenic phenotype were named tumor suppressor genes (TSGs). Allelic inactivation can take place as a single nucleotide variant (SNV), a short insertion or deletion (indel), an anomalous methylation or a copy number variant (CNV) [9]. Regarding their distribution, duplications are usually more abundant in healthy individuals than deletions because of their commonly milder phenotypic effect [10]. However, considering Knudson's hypothesis, a putative second hit would involve a deletion, thus leading to the somatic loss of heterozygosity (LOH).

Commonly used in recent years for the cost-effective discovery of pathogenic SNVs and indels, whole-exome sequencing (WES) has also marked a turning point for CNV and LOH detection. Despite the challenge of the uneven coverage distribution along the genome, WES approaches have emerged as a solid option for germline CNV calling [11], recently obtaining significant results in CRC predisposition [12]. Regarding tumor LOH detection, classic approaches were based on microsatellite markers around the gene of interest. However, ALFRED (allelic loss featuring rare damaging), a novel approach using WES data, has been recently developed to predict putative genes affected by LOH. It is a statistical method capable of inferring LOH status by testing for the allelic imbalance between germline and tumor sequencing data [13].

By means of the development of a combined germline–tumor WES analysis, the purpose of this study was the identification of novel candidate genes involved in germline predisposition to familial CRC. The potential TSG role of the selected candidates was assessed according to Knudson's two-hit hypothesis.

2. Results

2.1. Two-Hit Prioritization Strategy Identified New Candidate Genes for CRC Germline Predisposition

WES was performed in 18 unrelated familial CRC patients both in germline and tumor DNA. Prior to data analysis, quality control verifications were carried out. All germline samples yielded good results, with a mean coverage higher than 95× in all cases, resulting in approximately 4 gigabases sequenced per sample. However, two of the tumor samples (FAM22 and H461) showed a significant low value of shared exome regions sequenced (Figure S1) and were finally discarded.

An in-house pipeline was used to identify and filter genetic variants, including SNVs, indels, CNVs and LOHs, in germline and tumor WES data. Those rarest and potentially harmful variants with a function compatible with CRC susceptibility were highlighted. The prioritization strategy selected as candidates those genes affected by two hits according to Knudson's hypothesis and, therefore, those which are susceptible to have a TSG role.

Regarding germline CNVs, after their integrated calling using two different algorithms and frequency filtering, seven different rare variants were selected (five duplications and two deletions) (Table S1). However, functions and previously linked phenotypes of the affected genes were not sufficiently relevant to CRC, resulting in their not being further considered as putative germline mutational events. On the other hand, SNVs and indels recorded a total of 494 and 42 germline variants, respectively, after filtering. Thus, only the first hits in the form of SNVs and indels were finally taken under consideration, whereas second hits were selected from the whole spectrum of genetic alterations analyzed (SNV, indels and LOHs).

A total of 143 genes carried a germline–tumor pair of potentially disruptive variants in our samples. Among them, a germline SNV followed by a different somatic SNV was identified in *ADCY8*, *HSPG2* and *TTN*. No indel was found as a tumor second hit. The *TTN* gene encodes for a giant protein of more than 30,000 amino acids, thus having a higher probability of accumulating genetic alterations simply by chance. Considering also its function as a muscular protein, it was discarded as a potential cause for CRC predisposition. Therefore, *ADCY8* and *HSPG2* were selected as the most promising candidates from this double germline–tumor disrupting SNV approach (Table 1).

On the other hand, 141 genes were predicted to be affected by LOH as somatic mutational events with an SNV/indel as a germline first hit (133 SNVs and 8 indels) (Table S2). Interestingly, LOH was also predicted for *HSPG2* gene, thus presenting both kinds of second hits. Among the 141 germline–tumor pairs of potentially disruptive variants, we pursued an additional prioritization process to better select candidate genes with a plausible implication in CRC predisposition. In this regard, manual curation taking into account protein function compatible with CRC or cancer, as well as previously reported links with susceptibility to CRC or other neoplasms, was considered. A summary of the final 16 functionally prioritized candidates for germline SNV and tumor LOH prediction is shown in Table 2. Interestingly, DNA repair was one of most enriched functions among candidates, with 7 out of 16 genes (43.8%) linked to this cellular mechanism (*BRCA2*, *BLM*, *ERCC2*, *PARP2*, *RECQL*, *REV3L* and *RIF1*). It is also interesting to highlight candidate genes involved in hereditary cancer (*BRCA2*, *BLM*, *ERCC2*, *SMARCA4*) or connected to inherited CRC, such as Cowden syndrome (*SEC23B*) and Peutz–Jeghers syndrome (*STK11IP*). Taking this into account, 10 genes were selected as the best candidates for CRC germline predisposition from the approach of germline SNV/indel and somatic LOH including *BRCA2*, *BLM*, *ERCC2*, *PARP2*, *RECQL*, *REV3L*, *RIF1*, *SEC23B*, *SMARCA4* and *STK11IP*.

Table 1. Description of the genes carrying a potentially disruptive germline SNV (single nucleotide variant) and a different SNV in the matched-tumor sample.

Gene	Family	RefSeq Transcript	Hit	Genetic Variant	Path. Tools	DAMpred	ExAC Freq.	Protein Domain	Protein Function
ADCY8	FAMN4	NM_001115.2	1st	c.1747G>A p.(Glu583Lys)	5/6	−	21/60,697	Adenylyl cyclase class-3/4/guanylyl cyclase domain	Biosynthesis of cAMP from ATP
			2nd	c.458C>T p.(Ile153Thr)	4/6	+	0/60,706	Interaction with *ORAI1*, *STIM1*, *PPP2CA* and *PPP2R1A*	
HSPG2	FAM23	NM_005529.7	1st	c.3148G>A p.(Gly1050Ser)	3/6	−	3/60,456	Laminin IV type A domain	Component of vascular extracellular matrix, regulation of angiogenesis and cell growth
			2nd	c.7406C>T p.(Thr2469Met)	4/6	−	0/60,706	Immunoglobulin-like C2-type domain	

Abbreviations: DAMpred, disease-associated mutation prediction, affects protein structure (+), no effect on protein structure (−); ExAC, Exome Aggregation Consortium; Freq., frequency; Path., pathogenicity, cAMP: cyclic AMP.

Table 2. Candidate genes for germline colorectal cancer (CRC) predisposition selected after the two-hit prioritization strategy. In all cases, a first single nucleotide variant (SNV)/indel hit was present in the germline and a second loss of heterozygosity (LOH) hit was identified in the matched-tumor sample.

Gene	Family	RefSeq Transcript	Genetic Variant	Path. Tools	DAMpred	ExAC Freq.	Protein Domain	Protein Function
BRCA2	FAM20	NM_000059.3	c.4963delT p.(Tyr1655fs*15)	FS	n.a.	0/60,706	-	Double-strand break repair via homologous recombination, inherited predisposition to breast and ovarian cancer
BLM	FAMN4	NM_000057.4	c.2069C>T p.(Pro690Leu)	6/6	+	1/60,570	Helicase ATP-binding domain	DNA helicase, double-strand break repair via homologous recombination, regulation of cell cycle and apoptosis, DNA replication, telomere maintenance
ERCC2	H458	NM_000400.3	c.688G>A p.(Val230Ile)	4/6	–	0/60,706	Helicase ATP-binding domain	DNA helicase, transcription-coupled nucleotide excision repair, regulation of cell cycle
FAT2	FAMN3	NM_001447.2	c.1643T>C p.(Val548Ala)	5/6	–	0/60,706	Cadherin domain	Regulation of cell proliferation, cell adhesion
IGF2R	H466	NM_000876.3	c.232G>A p.(Gly78Arg)	6/6	+	1/60,684	-	Positive regulation of apoptosis
LATS2	H460	NM_014572.3	c.337G>A p.(Asp113Asn)	5/6	–	1/56,138	Ubiquitin-associated domain	Positive regulation of apoptosis, regulation of cell cycle
PARP2	FAM20	NM_005484.3	c.910G>C p.(Glu304Gln)	3/6	–	3/60,208	Poly(ADP-ribose) polymerase (PARP) alpha-helical domain	Base excision repair, extrinsic apoptotic signaling pathway
PSMD9	H469	NM_002813.6	c.361A>T p.(Ser121Cys)	3/6	–	30/60,148	PDZ domain	Subunit of 26S proteasome, regulation of apoptosis and cell cycle, regulation of ubiquitin-protein ligase activity
RASSF6	H460	NM_201431.2	c.779C>T p.(Pro260Leu)	6/6	–	53/60,475	Ras-associating domain	Positive regulation of apoptosis
RECQL	H466	NM_002907.4	c.221_225delinsAATGT p.(Pro74_Trp75delinsGlnCys)	6/6	+	0/60,706	-	DNA helicase, double-strand break repair via homologous recombination, DNA replication
RERGL	H466	NM_024730.3	c.362T>C p.(Val121Ala)	6/6	+	54/60,446	-	Unknown (closely related to *RERG*, which functions as a negative regulator of cell growth [14])
REV3L	FAM3	NM_002912.4	c.559A>T p.(Arg187Trp)	5/6	–	0/60,706	Exonuclease domain (family B of DNA polymerases)	DNA repair, translesion DNA synthesis
RIF1	H460	NM_018151.4	c.4262G>A p.(Arg1421His)	4/6	+	5/59,938	-	Double-strand break repair via nonhomologous end joining, telomere maintenance
SEC23B	H470	NM_032985.5	c.531G>C p.(Glu177Asp)	4/6	–	1/60,706	Sec23/Sec24 trunk domain	Intracellular protein transport, associated with inherited cancer predisposition Cowden Syndrome
SMARCA4	FAM3	NM_003072.3	c.295C>T p.(Arg99Trp)	5/6	–	1/60,196	-	Regulation of cell growth, regulation of cell cycle, chromatin remodeling
STK11IP	H470	NM_052902.4	c.1214C>T p.(Pro405Leu)	5/6	–	51/59,930	-	Interaction with *STK11* (serine/threonine kinase activity, negative regulation of cell growth, Peutz-Jeghers CRC predisposition syndrome)

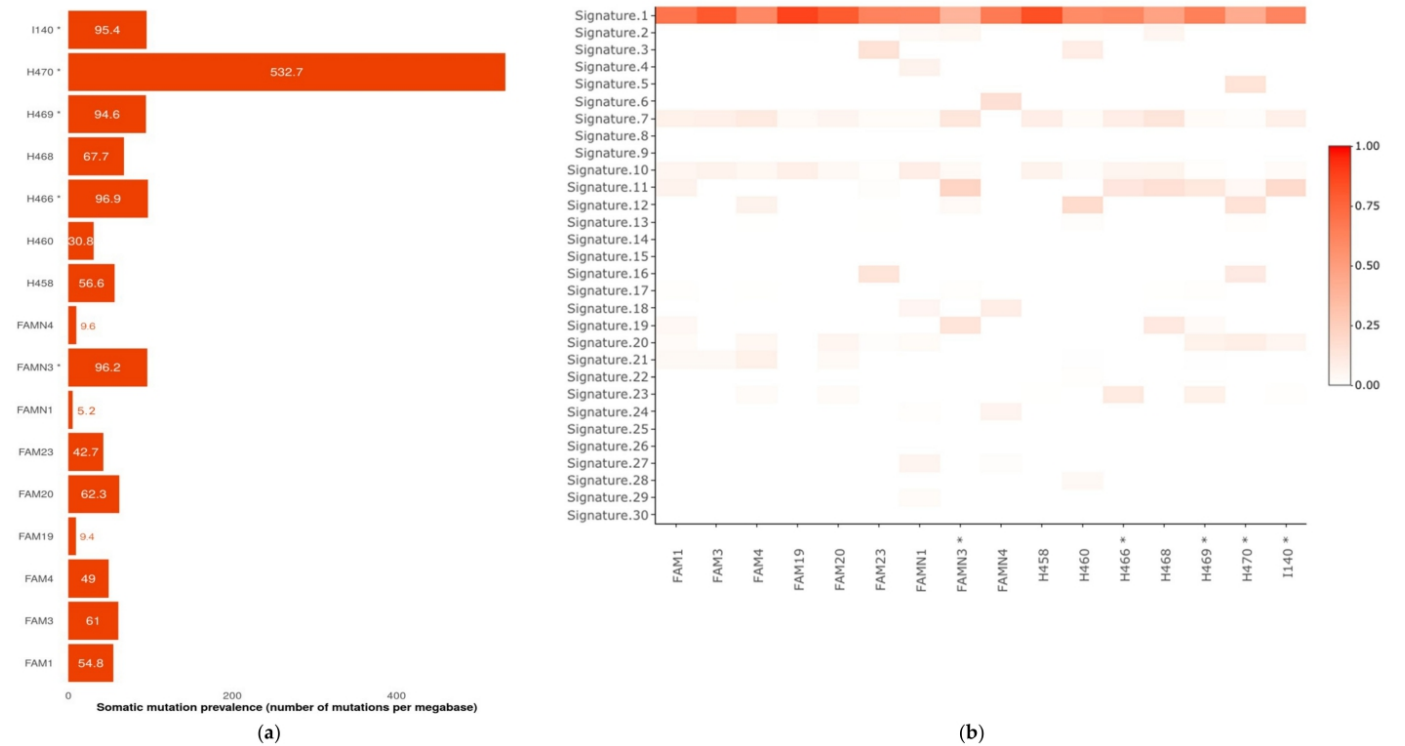
Abbreviations: DAMpred: disease-associated mutation prediction, affects protein structure (+), no effect on protein structure (−); n.a., not available; ExAC, Exome Aggregation Consortium; Freq., frequency; FS, frameshift; Path., pathogenicity.

All variants located in the 12 final candidate genes were validated by the manual inspection of WES data. In addition, a case-control enrichment analysis for the 12 final candidate genes was performed using a publicly available independent cohort of 1006 patients of familial early-onset CRC (CanVar) and the Exome Aggregation Consortium (ExAC) database. We checked if rare and potentially pathogenic variants in the 12 final candidate genes were also present in this CRC cohort and tested if they were more frequent than in the ExAC control dataset. Potentially pathogenic and rare variants were found in CanVar for all 12 genes assessed. *ADCY8*, *BLM*, *BRCA2*, *ERCC2*, *REV3L*, *RIF1*, *SEC23B*, *SMARCA4* and *STK11IP* were highlighted for harboring a significant enrichment in CRC cases for more than 50% of the potentially disrupting variants (Table S3).

2.2. Somatic Mutational Profiling Detected Hypermutated Tumors Compatible with A Germline Defect Etiology

Different somatic specific features were assessed in order to identify possible links with germline CRC predisposition that could help in the selection of the most suitable candidate genes. Tumor mutational burden analysis presented a large number of mutations per sample, with 5 out of 16 samples showing a hypermutated profile with more than 90 mutations per megabase (Mb), and a median of 58.8 mutations per Mb in the whole cohort (Figure 1). One of the hypermutated samples, H466 (96.9 mutations per Mb), was affected by the putative loss of function of a DNA repair-associated gene according to the two-hit prioritization strategy, *RECQL*. A germline deficiency in the DNA repair pathways affected by this gene could explain both the inherited predisposition to CRC and the elevated tumor mutational prevalence shown by the patient. An ultrahypermutated sample with more than 500 mutations per Mb (sample H470) was also identified. Interestingly, no deleterious mutation in *POLE*, *POLD1* or the MMR genes was found in the germline or somatic profile of this patient.

Regarding mutational signatures, the typical profile of a microsatellite-stable and *POLE*-wild-type CRC is shown by the mutational profile reconstruction using the 30 reference signatures of the COSMIC database (Catalogue of Somatic Mutations in Cancer; https://cancer.sanger.ac.uk/cosmic/signatures). This included a strong predominance of clock-like mutational signature 1, directly associated with the age of onset, along with a low prevalence of signatures related with MMR deficiency (signatures 6, 15, 20 and 26) and *POLE* mutations (signature 10). Specifically, signature 1 has been linked with the spontaneous deamination of 5-methylcytosine at NpCpG trinucleotides leading to T/G mismatches which are not repaired before DNA replication and, therefore, predominantly generate C>T mutations. Interestingly, none of the other signatures currently associated with a particular deficiency in a DNA repair pathway (signatures 2 and 13 with APOBEC activity, signature 3 with double-strand break repair via homologous recombination and signatures 18 and 30 with base excision repair) were detected as a relevant contributor in any of the analyzed tumor samples. Mutational signatures 7 and 11 were the other two signatures with a greater prevalence in our cohort, although they contributed just 6% to the profile reconstruction on average. A link between UV light exposure and signature 7 has been consistently demonstrated, whereas signature 11 has been mostly associated with alkylating chemotherapy treatments. Both etiologies were not apparently relevant for CRC germline predisposition.



(**a**) (**b**)

Figure 1. Somatic mutational profile analysis performed with the Mutational Signatures in Cancer (MuSiCa) tool in 16 germline–tumor paired samples. (**a**) Mutational prevalence (number of mutations per sequenced Mb). Hypermutated samples (≥90 mutations/Mb) are marked with an asterisk (*); (**b**) mutational signature refitting analysis showing the contributions of the 30 Catalogue of Somatic Mutations in Cancer (COSMIC) reference mutational signatures in the mutational catalogues of the samples of the study.

3. Discussion

An integrated germline–tumor WES analysis was performed in 16 unrelated samples after quality control filtering, resulting in the prioritization of 12 new candidate genes for CRC germline predisposition. A germline SNV and a tumor SNV were identified in *ADCY8* and *HSPG2* genes, whereas a germline SNV/indel and somatic LOH of the wild-type allele was predicted for *BRCA2*, *BLM*, *ERCC2*, *PARP2*, *RECQL*, *REV3L*, *RIF1*, *SEC23B*, *SMARCA4* and *STK11IP*.

ADCY8 is a membrane-bound enzyme that catalyzes the formation of cyclic AMP (cAMP) from ATP. The cAMP pathway was already found to be associated with cancer, with the overexpression of *ADCY3* increasing oncogenic potential in gastric cancer cells [15] and *ADCY8* acting itself as a risk modifier in glioma [16]. *HSPG2* encodes for the perlecan protein, an essential extracellular matrix component. Its effect on CRC was described using cell lines and tumor xenografts and allografts, where an oncogene role promoting tumor growth and angiogenesis was found [17]. Thus, both genes were not in accordance with the TSG role expected for the genes prioritized by our integrated germline–tumor analysis and were therefore discarded as the putative cause of the inherited predisposition to CRC in the affected families.

A role in DNA repair, along with a previous association with hereditary cancer syndromes, drove the prioritization of the germline SNV/indel and somatic LOH candidates. *RECQL* presented both alterations in a patient with a hypermutated tumor, thus suggesting the hypothesis of a deregulation of a DNA repair mechanism causing a rapid increase in the number of tumor mutations. In this case, the *RECQL* variant (p.Pro74_Trp75delinsGlnCys) was not reported in ExAC and had a potential disruptive effect in the protein structure (Table 2). This gene encodes for a DNA helicase belonging to the RecQ family, responsible for the unwinding of double-stranded DNA and therefore implicated in both DNA replication and repair [18]. Thus, the loss of function of *RECQL* would affect the maintenance of chromosomal stability. In this regard, mutations in this gene have already been linked to breast cancer predisposition [19]. Interestingly, other key members of the same protein family have been associated with well-known recessive cancer predisposition syndromes (*BLM*, Bloom syndrome; *RECQL4*, Rothmund–Thompson syndrome; *WRN*, Werner syndrome) [20]. *BLM* was also found to be mutated in the germline DNA of one patient in our cohort and prioritized by our two-hit integrated analysis. However, although the missense mutation found (p.Pro690Leu) was predicted to be deleterious by different in silico tools and located at the helicase domain of the protein, the tumor showed a low mutation burden. This could indicate either a non-significant effect of the identified variant in *BLM* function, or an association with a distinct carcinogenic mechanism, such us chromosomal instability (linked to a high number of CNVs and aneuploidy instead of SNVs). Interestingly, our study highlights the link between CRC and breast cancer predisposition genes, as well as the relevance of the Fanconi anemia pathway, as also underlined by previous studies [21–23].

BRCA2 and *ERCC2* are also linked with classical cancer predisposition syndromes, hereditary breast and ovarian cancer (HBOC) and xeroderma pigmentosum (XP), respectively [20]. In the case of *BRCA2*, the germline frameshift variant found in family FAM20 (p.Tyr1655fsTer15) was classified as pathogenic in ClinVar for HBOC. A role for this variant in the CRC predisposition was also suggested in a previous study using the same cohort [21] and additionally supported by the presence of additional breast cancer patients in the family (Figure S2). Accordingly, the strength of this association discarded the other prioritized gene in the family, *PARP2*, which was also implicated in DNA repair. In addition, *BRCA2* mutations were found to be significantly enriched in the case-control analysis but not *PARP2* mutations. On the other hand, *ERCC2* encodes for a subunit of the DNA helicase in charge of the nucleotide excision repair (NER) mechanism [24]. Homozygous or compound heterozygous mutations in this gene are known to cause XP, a condition responsible for skin cancer predisposition [25]. In a recent study, its association with breast and ovarian cancer susceptibility was also proposed [26]. Interestingly, a specific mutational signature characterized by a broad distribution of nucleotide changes have recently been associated with somatic mutations in *ERCC2* [27]. However, in the somatic analysis performed for the patient harboring germline mutation in this gene (H458), this signature was

not identified. In contrast, a strong predominance of age-related signature 1 was found (84% of somatic mutations explained by this mutational source), along with a small contribution of signature 7 (9%) (Figure 1). UV-derived mutations, commonly responsible for the latter signature, are repaired by NER, potentially altered in this case by the *ERCC2* inactivation and thus explain this specific contribution to the somatic mutational profile observed. The germline mutation detected in our study (p.V230I) is affecting the helicase ATP binding domain of the protein and has not been detected in the ExAC database, thus suggesting its potential disruptive effect. In addition, disruptive variants in this gene were found to be significantly enriched in the case-control analysis performed in familial early-onset CRC patients.

REV3L and *RIF1* were also prioritized by our integrated analysis and involved in translesion DNA synthesis and nonhomologous end-joining DNA repair mechanisms, respectively [28–30]. Both carried potentially pathogenic germline alterations according to the different evidence assessed (Table 2), whereas the corresponding tumors showed a moderately mutated profile (61 and 30.8 mutations per Mb, respectively). In addition, disruptive variants in both genes were found to be more significantly enriched in cases than controls. *REV3L* was prioritized in family FAM3, where also a double inactivation of *SMARCA4* was predicted by our integrated analysis. The somatic LOH status of both alterations were validated for this specific family using Sanger sequencing in previous studies [21,31]. The results did not confirm an LOH of the wild-type allele in the case of *SMARCA4*, whereas it was detected for *REV3L*, thus supporting this gene as a better candidate.

An ultrahypermutated tumor was also found in one patient of our cohort (H470). The high number of somatic mutations detected cannot be explained by classic somatic hypermutation drivers (*POLE*, *POLD1* and the MMR genes) [32], thus suggesting a specific alteration of another DNA repair mechanism responsible for the phenotype. Interestingly, no gene implicated in this cellular mechanism was identified by our integrated analysis. In contrast, *SEC23B* and *STK11IP* were the genes prioritized through our approach for this patient. The specific functions of proteins encoded by these genes are not directly related with CRC, although both are associated to cancer predisposition syndromes. *SEC23B* is implicated in endoplasmic reticulum to Golgi apparatus transport [33], and has also been recently associated with Cowden syndrome [34]. This inherited condition is linked to hamartomatous polyps and elevated susceptibility to different epithelial cancers, being caused by germline mutations in *PTEN* in most cases [34,35]. On the other hand, Peutz–Jeghers syndrome is an autosomal dominant CRC predisposition syndrome also related to hamartomas and is mainly caused by germline mutations in the TSG *STK11* [5]. *STK11IP*, whose function is not currently broadly described, is known to be interacting with *STK11*, and therefore potentially implicated in CRC predisposition [36].

Our development of a germline–tumor prioritization strategy was in accordance with recent recommendations from the Germline/Somatic Variant Subcommittee (GSVS) of the Clinical Genome Resource (ClinGen), on the use of tumor sequencing data for germline variant interpretation [37]. Even if the loss of heterozygosity and second mutation of the alternative allele assessment were not directly recommended for clinical routine, both pieces of evidence supporting the Knudson's two-hit hypothesis could add a great value in the variant prioritization process in a comprehensive germline–tumor WES study. In fact, the power of this approach have been proven by previous studies using a similar methodology based in two-hit hypothesis assessment [13,38–41]. In addition, both tumor phenotypic features analyzed, tumor mutational burden and signatures, were recommended to improve the support of the pathogenicity of germline variants by this and additional studies [37,42]. However, no methylation data may impact the assessment of the two-hit hypothesis, missing those genes affected by epigenetic silencing. In any case, further functional studies and replication in additional cohorts will be needed in order to further confirm the identified potential candidates for CRC germline predisposition.

4. Materials and Methods

4.1. Patients

Eighteen unrelated Spanish patients (one per family) with unaffiliated strong CRC aggregation compatible with an autosomal dominant pattern of inheritance and available germline and tumor DNA samples were selected from a previously described cohort of 71 individuals from 38 families (Figure S2). Families were selected based on the following criteria: three or more relatives with CRC, two or more consecutive affected generations and at least one CRC diagnosed before the age of 60. The entire cohort had germline WES data available from previous studies [12,21,31]. The presence of germline alterations in well-known genes related with hereditary CRC syndromes (*APC*, *MUTYH* and the DNA MMR genes) were previously discarded for all probands. The present study was approved by the Institutional Ethics Committee (register number 2011/6440, date of approval 22/03/2011). Written informed consent was obtained in all cases.

Matched tumor DNA samples were used to perform WES when available with an optimal quantity and quality from our cohort of 38 CRC families. Tumor DNA was isolated from formalin-fixed paraffin-embedded tissue using the QIAamp Tissue Kit (Qiagen, Redwood City, CA, USA) following the manufacturer's instructions and reaching a percentage of tumor cells of 70–80% among all 18 available samples. Germline DNA samples of other members of the family diagnosed with CRC, advanced adenoma (i.e., lesion size $\geq$ 1 cm, villous architecture or high-grade dysplasia) or other extracolonic cancers were also used in previous studies for germline segregation.

4.2. Whole Exome Sequencing

Germline WES data were available from previous studies [12,21,31]. WES was performed in tumor samples of selected patients using the HiSeq2000 platform (Illumina, San Diego, CA, USA) and SureSelectXT Human All Exon v5 kit (Agilent, Santa Clara, CA, USA) for exon enrichment. Indexed libraries were pooled and massively parallel-sequenced using a paired-end 2 $\times$ 75 bp read length protocol.

The quality control of sequencing data was made in all samples previous to their analysis using the Real-Time Analysis software sequence pipeline (Illumina). Additionally, the proportion of all shared exome regions sequenced with a coverage $\geq 10\times$ was evaluated for tumor samples. A good ratio of shared regions with high coverage ($\geq$ 70%) was expected in good-quality samples, whereas low-quality ones were characterized by a significant drop in this percentage.

WES data analysis was performed in accordance with the workflow displayed in Figure 2. The Burrows–Wheeler Aligner (BWA-MEM algorithm) was used for read mapping to the human reference genome (build hs37d5, based on NCBI GRCh37) [43]. PCR duplicates were discarded using the MarkDuplicates tool from Picard, and then indel realignment and base quality score recalibration were performed with the Genome Analysis Toolkit (GATK, Broad Institute, Cambridge, USA) [44].

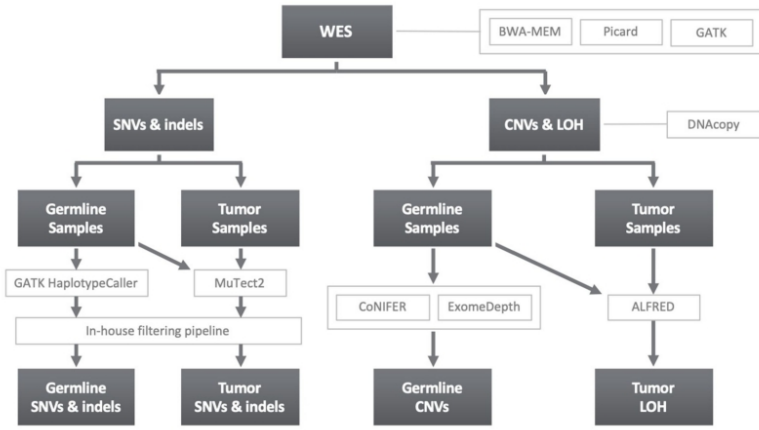


Figure 2. Methodology schematic for variant identification, showing the software used in each analysis step for the different classes of genetic variation considered. WES, whole-exome sequencing; BWA-MEM, Burrows-Wheeler Alignment Tool; GATK, Genome Analysis Toolkit; SNV, single nucleotide variant; indels, insertion and deletion variants; CNV, copy number variant; LOH, loss of heterozygosity; MuTect2, somatic SNV and indel variants caller.

4.3. Variant Calling and Filtering

4.3.1. SNVs and Indels

The GATK tools HaplotypeCaller and MuTect2 were used for SNV and short indels calling for germline and tumor samples, respectively [44]. To improve germline variant filtering with MuTect2, a panel of 71 available germline CRC samples from the whole cohort was used in the case of five of the tumor samples, whereas an in-house pipeline from the CNAG-CRG (Centre Nacional d'Anàlisi Genòmica-Centre de Regulació Genòmica, Barcelona, Spain) was implemented for the rest. Regarding variant annotation, different databases were considered, including SnpEff, ANNOVAR and dbNSFP for pathogenicity and variant position annotation. PhyloP (phyloP46way_placental score $\geq$1.6), SIFT (prediction of damaging), PolyPhen2 (HumVar prediction of probably damaging or possibly damaging), MutationTaster (prediction of disease-causing or disease-causing-automatic), LRT (prediction of deleterious) and CADD (Phred score $\geq$15) were used for the pathogenicity prediction of missense variants. Germline WES data was analyzed through an in-house R language pipeline described in previous studies [12,21,31]. Functions related with CRC or cancer in general were prioritized. DNA repair, apoptosis, autophagy, cell growth, cell proliferation, inflammatory response, cell cycle, angiogenesis, cell differentiation, cell adhesion and chromatin modification, among others, were included. Concerning tumor SNVs and indels, a similar filtering pipeline was used, restraining selected variants to those having a coverage $\geq$10$\times$ both in germline and somatic samples, an alternative allelic frequency in the tumor $\geq$20%, and also selecting truncating or missense variants fulfilling at least three of the missense pathogenicity tools criteria.

4.3.2. Copy Number Variants and Loss of Heterozygosity

The DNAcopy R package was used for the implementation of the circular binary segmentation algorithm [45]. This was required for the fragmentation of the WES data in order to identify genomic regions with an abnormal value of copy number. After segmentation, CoNIFER and Exome Depth were used in germline data for CNV identification as previously described [12], whereas ALFRED was used to predict the LOH of the wild-type allele in tumor samples [13].

4.4. Variant Prioritization and Validation

After the automatic filtering process was performed for all variant types considered, a large number of potentially pathogenic alterations were identified for every sample. Thus, an additional prioritization process was required in order to select those actually relevant for the phenotype under study. Taking advantage of the access to both germline and somatic WES data, an integrated strategy based on Knudson's two-hit hypothesis was developed in order to look for potential TSGs associated with CRC germline predisposition. Genes with a deleterious germline variant (first hit, SNV/indel or CNV) and a second mutational event in the tumor (second hit, SNV/indel or LOH) were thus prioritized.

The prioritization process was completed with an additional stringent functional selection of the candidate genes compatible with the TSG model expected. The most interesting final candidates were manually curated according to functional evidence. In addition, the amino acidic position of the variants within specific functional protein domains was checked using UniProtKB (http://www.uniprot.org/) and InterPro (http://www.ebi.ac.uk/interpro/), as well as a possible 3D protein structure destabilization effect by using the DAMpred tool (disease-associated mutation prediction; https://zhanglab.ccmb.med.umich.edu/DAMpred/). Special attention was paid to genes previously involved in predisposition to CRC and other neoplasms by reviewing the data present in OMIM (Online Mendelian Inheritance in Man; http://www.omim.org/) and ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/).

The final prioritized variants were validated by manual inspection of the WES data with the Integrative Genomics Viewer [46]. This high-performance data visualizer permits the exclusion of any possible sequencing artifacts, especially those due to strand bias. This is the case when the genotype information given by the data from the forward strand and the reverse strand is significantly different [47]. The CanVar browser [48], a resource of variant level frequency data from cancer germline sequencing studies containing 1006 familial early onset CRC patients, was also used to search for additional variants in this independent familial CRC cohort. Only rare variants (ExAC allele frequency $< 0.1\%$) and potentially pathogenic (CADD Phred score > 15) were considered. Variant enrichment was calculated by comparing the number of cases in the CanVar cohort with the number of controls in the ExAC repository using a Fisher's exact test.

4.5. Mutational Profiling and Mutational Signature Analysis

Somatic WES data was also specifically analyzed in order to look for particular tumor features supporting a hypothesis for the inherited predisposition to familial CRC in the samples considered. In this regard, both the tumor mutational burden and mutational signatures were taken into account. The MuSiCa (Mutational Signatures in Cancer) web application was used to perform these analyses [49]. The prevalence of somatic mutations was described as the total number of SNVs per Mb accumulated in a specific sample, assuming that an average WES sample accounts for 30 Mb with acceptable sequencing quality values. With respect to mutational signatures, the original computational framework described by Alexandrov and collaborators was considered [50,51]. Original mutational profiles of the analyzed samples were reconstructed by the non-negative least squares algorithm using the 30 reference signatures described in the COSMIC database [52].

5. Conclusions

Our integrated germline–tumor analysis based on Knudson's hypothesis allowed the identification of new potential genes implicated in the inherited predisposition to CRC. *BRCA2*, *BLM*, *ERCC2*, *RECQL*, *REV3L* and *RIF1* were among the most promising candidates, with some of them previously associated with predisposition syndromes to other cancers. DNA repair was found to be enriched among the genes prioritized by our approach, thus highlighting the importance of this cellular mechanism in germline predisposition to colorectal carcinogenesis.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/11/3/362/s1, Figure S1: histogram representing the percentage of genomic regions with a high-quality value of coverage (≥10×) with respect to all shared sequenced regions for each of the germline–tumor paired samples, Figure S2: pedigrees of the 18 families included in the study, Table S1: description of germline copy number variants after calling by CoNIFER and ExomeDepth, Table S2: description of genes and germline variants of the cases where a potentially pathogenic germline SNV/indel and tumor LOH were identified, Table S3: list of potentially pathogenic rare germline variants found in a cohort of 1006 familial early onset CRC patients corresponding to CanVar database.

Author Contributions: Conceptualization, M.D.-G., S.F.-E., C.E.-J. and S.C.-B.; Funding acquisition, S.C.-B. and A.C.; Investigation, M.D.-G., S.F.-E., S.P., F.S., J.M., C.A.-C., L.B. (Laia Bonjoch), A.G.-M., P.A.S.-R., C.E.-J. and S.C.-B.; Resources, T.O., M.C., J.J.L., EPICOLON consortium, A.C., L.B. (Luis Bujanda), J.C., F.B. and S.C.-B.; Software, M.D.-G., S.F.-E., S.P., F.S., M.V.-C., J.J.L., G.P., S.L. and S.B.; Supervision, S.C.-B., Visualization, M.D.-G., S.F.-E. and M.V.-C.; Writing—original draft, M.D.-G. and S.C.-B.; Writing—review & editing, M.D.-G., S.F.-E., S.P., F.S., J.M., C.A.-C., L.B. (Laia Bonjoch), A.G.-M., P.A.S.-R., C.E.-J., T.O., M.C., M.V.-C., J.J.L., G.P., S.L., S.B., A.C., L.B. (Luis Bujanda), J.C., F.B. and S.C.-B.

Funding: M.D.-G. was supported by a contract from Agència de Gestió d'Ajuts Universitaris i de Recerca -AGAUR- (Generalitat de Catalunya, 2018FI_B1_00213). S.F.-E., J.M., C.A.-C., C.E.-J. and J.J.L. were supported by a contract from CIBEREHD. CIBEREHD is funded by the Instituto de Salud Carlos III. This research was supported by grants from Fondo de Investigación Sanitaria/FEDER (17/00878), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), PERIS (SLT002/16/00398, Generalitat de Catalunya), CERCA Programme (Generalitat de Catalunya) and Agència de Gestió d'Ajuts Universitaris i de Recerca (Generalitat de Catalunya, GRPRE 2017SGR21, GRC 2017SGR653). This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology). www.cost.eu\T1\textquoteright.

Acknowledgments: We are sincerely grateful to the patients, Baldo Oliva, CNAG, the Biobank of Hospital Clínic–IDIBAPS and Biobanco Vasco. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Mathers, C.; Parkin, D.M.; Piñeros, M.; Znaor, A.; Bray, F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **2019**, *144*, 1941–1953. [CrossRef] [PubMed]
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
3. Wei, E.K.; Colditz, G.A.; Giovannucci, E.L.; Wu, K.; Glynn, R.J.; Fuchs, C.S.; Stampfer, M.; Willett, W.; Ogino, S.; Rosner, B. A Comprehensive Model of Colorectal Cancer by Risk Factor Status and Subsite Using Data From the Nurses' Health Study. *Am. J. Epidemiol.* **2017**, *185*, 224–237. [CrossRef]
4. Lichtenstein, P.; Holm, N.V.; Verkasalo, P.K.; Iliadou, A.; Kaprio, J.; Koskenvuo, M.; Pukkala, E.; Skytthe, A.; Hemminki, K. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **2000**, *343*, 78–85. [CrossRef] [PubMed]
5. Tomlinson, I. The Mendelian colorectal cancer syndromes. *Ann. Clin. Biochem. Int. J. Biochem. Lab. Med.* **2015**, *53*, 690–692. [CrossRef] [PubMed]
6. Jasperson, K.W.; Tuohy, T.M.; Neklason, D.W.; Burt, R.W. Hereditary and familial colon cancer. *Gastroenterology* **2010**, *138*, 2044–2058. [CrossRef] [PubMed]
7. Valle, L. Recent Discoveries in the Genetics of Familial Colorectal Cancer and Polyposis. *Clin. Gastroenterol. Hepatol.* **2017**, *15*, 809–819. [CrossRef]
8. Knudson, A.G. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 820–823. [CrossRef] [PubMed]
9. Carvalho, C.M.B.; Lupski, J.R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **2016**, *17*, 224–238. [CrossRef]
10. Zarrei, M.; MacDonald, J.R.; Merico, D.; Scherer, S.W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **2015**, *16*, 172–183. [CrossRef]
11. Tan, R.; Wang, Y.; Kleinstein, S.E.; Liu, Y.; Zhu, X.; Guo, H.; Jiang, Q.; Allen, A.S.; Zhu, M. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum. Mutat.* **2014**, *35*, 899–907. [CrossRef] [PubMed]

12. Franch-Expósito, S.; Esteban-Jurado, C.; Garre, P.; Quintanilla, I.; Duran-Sanchon, S.; Díaz-Gay, M.; Bonjoch, L.; Cuatrecasas, M.; Samper, E.; Muñoz, J.; et al. Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis. *J. Genet. Genom.* **2018**, *45*, 41–45. [CrossRef]
13. Park, S.; Supek, F.; Lehner, B. Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. *Nat. Commun.* **2018**, *9*, 2601. [CrossRef] [PubMed]
14. Finlin, B.S.; Gau, C.-L.; Murphy, G.A.; Shao, H.; Kimel, T.; Seitz, R.S.; Chiu, Y.-F.; Botstein, D.; Brown, P.O.; Der, C.J.; et al. RERG Is a Novel ras-related, Estrogen-regulated and Growth-inhibitory Gene in Breast Cancer. *J. Biol. Chem.* **2001**, *276*, 42259–42267. [CrossRef] [PubMed]
15. Hong, S.-H.; Goh, S.-H.; Lee, S.J.; Hwang, J.-A.; Lee, J.; Choi, I.-J.; Seo, H.; Park, J.-H.; Suzuki, H.; Yamamoto, E.; et al. Upregulation of adenylate cyclase 3 (ADCY3) increases the tumorigenic potential of cells by activating the CREB pathway. *Oncotarget* **2013**, *4*, 1791–1803. [CrossRef] [PubMed]
16. Warrington, N.M.; Sun, T.; Luo, J.; McKinstry, R.C.; Parkin, P.C.; Ganzhorn, S.; Spoljaric, D.; Albers, A.C.; Merkelson, A.; Stewart, D.R.; et al. The Cyclic AMP Pathway Is a Sex-Specific Modifier of Glioma Risk in Type I Neurofibromatosis Patients. *Cancer Res.* **2015**, *75*, 16–21. [CrossRef] [PubMed]
17. Sharma, B.; Handler, M.; Eichstetter, I.; Whitelock, J.M.; Nugent, M.A.; Iozzo, R. V Antisense targeting of perlecan blocks tumor growth and angiogenesis in vivo. *J. Clin. Investig.* **1998**, *102*, 1599–1608. [CrossRef]
18. Sharma, S.; Doherty, K.M.; Brosh, R.M. Mechanisms of RecQ helicases in pathways of DNA metabolism and maintenance of genomic stability. *Biochem. J.* **2006**, *398*, 319–337. [CrossRef]
19. Cybulski, C.; Carrot-Zhang, J.; Kluźniak, W.; Rivera, B.; Kashyap, A.; Wokołorczyk, D.; Giroux, S.; Nadaf, J.; Hamel, N.; Zhang, S.; et al. Germline RECQL mutations are associated with breast cancer susceptibility. *Nat. Genet.* **2015**, *47*, 643. [CrossRef]
20. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **2014**, *505*, 302–308. [CrossRef]
21. Esteban-Jurado, C.; Franch-Expósito, S.; Muñoz, J.; Ocaña, T.; Carballal, S.; López-Cerón, M.; Cuatrecasas, M.; Vila-Casadesús, M.; Lozano, J.J.; Serra, E.; et al. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *Eur. J. Hum. Genet.* **2016**, *24*, 1501–1505. [CrossRef] [PubMed]
22. García, M.J.; Fernández, V.; Osorio, A.; Barroso, A.; Llort, G.; Lázaro, C.; Blanco, I.; Caldés, T.; de la Hoya, M.; Ramón y Cajal, T.; et al. Analysis of FANCB and FANCN/PALB2 fanconi anemia genes in BRCA1/2-negative Spanish breast cancer families. *Breast Cancer Res. Treat.* **2009**, *113*, 545–551. [CrossRef] [PubMed]
23. Tedaldi, G.; Tebaldi, M.; Zampiga, V.; Danesi, R.; Arcangeli, V.; Ravegnani, M.; Cangini, I.; Pirini, F.; Petracci, E.; Rocca, A.; et al. Multiple-gene panel analysis in a case series of 255 women with hereditary breast and ovarian cancer. *Oncotarget* **2017**, *8*, 47064–47075. [CrossRef] [PubMed]
24. Coin, F.; Marinoni, J.-C.; Rodolfo, C.; Fribourg, S.; Pedrini, A.M.; Egly, J.-M. Mutations in the XPD helicase gene result in XP and TTD phenotypes, preventing interaction between XPD and the p44 subunit of TFIIH. *Nat. Genet.* **1998**, *20*, 184. [CrossRef] [PubMed]
25. Frederick, G.D.; Amirkhan, R.H.; Schultz, R.A.; Friedberg, E.C. Structural and mutational analysis of the xeroderma pigmentosum group D (XPD) gene. *Hum. Mol. Genet.* **1994**, *3*, 1783–1788. [CrossRef] [PubMed]
26. Rump, A.; Benet-Pages, A.; Schubert, S.; Kuhlmann, J.D.; Janavičius, R.; Macháčková, E.; Foretová, L.; Kleibl, Z.; Lhota, F.; Zemankova, P.; et al. Identification and Functional Testing of ERCC2 Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer. *PLOS Genet.* **2016**, *12*, e1006248. [CrossRef] [PubMed]
27. Kim, J.; Mouw, K.W.; Polak, P.; Braunstein, L.Z.; Kamburov, A.; Tiao, G.; Kwiatkowski, D.J.; Rosenberg, J.E.; Van Allen, E.M.; D'Andrea, A.D.; et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **2016**, *48*, 600–606. [CrossRef]
28. Yang, L.; Shi, T.; Liu, F.; Ren, C.; Wang, Z.; Li, Y.; Tu, X.; Yang, G.; Cheng, X. REV3L, a Promising Target in Regulating the Chemosensitivity of Cervical Cancer Cells. *PLoS ONE* **2015**, *10*, e0120334. [CrossRef]
29. Chapman, J.R.; Barral, P.; Vannier, J.-B.; Borel, V.; Steger, M.; Tomas-Loba, A.; Sartori, A.A.; Adams, I.R.; Batista, F.D.; Boulton, S.J. RIF1 Is Essential for 53BP1-Dependent Nonhomologous End Joining and Suppression of DNA Double-Strand Break Resection. *Mol. Cell* **2013**, *49*, 858–871. [CrossRef]
30. Escribano-Díaz, C.; Orthwein, A.; Fradet-Turcotte, A.; Xing, M.; Young, J.T.F.; Tkáč, J.; Cook, M.A.; Rosebrock, A.P.; Munro, M.; Canny, M.D.; et al. A Cell Cycle-Dependent Regulatory Circuit Composed of 53BP1-RIF1 and BRCA1-CtIP Controls DNA Repair Pathway Choice. *Mol. Cell* **2013**, *49*, 872–883. [CrossRef]

31. Esteban-Jurado, C.; Vila-Casadesús, M.; Garre, P.; Lozano, J.J.; Pristoupilova, A.; Beltran, S.; Muñoz, J.; Ocaña, T.; Balaguer, F.; López-Cerón, M.; et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet. Med.* **2015**, *17*, 131–142. [CrossRef] [PubMed]
32. Campbell, B.B.; Light, N.; Fabrizio, D.; Zatzman, M.; Fuligni, F.; de Borja, R.; Davidson, S.; Edwards, M.; Elvin, J.A.; Hodel, K.P.; et al. Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **2017**, *171*, 1042–1056e10. [CrossRef] [PubMed]
33. Schwarz, K.; Iolascon, A.; Verissimo, F.; Trede, N.S.; Horsley, W.; Chen, W.; Paw, B.H.; Hopfner, K.-P.; Holzmann, K.; Russo, R.; et al. Mutations affecting the secretory COPII coat component SEC23B cause congenital dyserythropoietic anemia type II. *Nat. Genet.* **2009**, *41*, 936. [CrossRef] [PubMed]
34. Yehia, L.; Niazi, F.; Ni, Y.; Ngeow, J.; Sankunny, M.; Liu, Z.; Wei, W.; Mester, J.L.; Keri, R.A.; Zhang, B.; et al. Germline Heterozygous Variants in SEC23B Are Associated with Cowden Syndrome and Enriched in Apparently Sporadic Thyroid Cancer. *Am. J. Hum. Genet.* **2015**, *97*, 661–676. [CrossRef] [PubMed]
35. Liaw, D.; Marsh, D.J.; Li, J.; Dahia, P.L.M.; Wang, S.I.; Zheng, Z.; Bose, S.; Call, K.M.; Tsou, H.C.; Peacoke, M.; et al. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.* **1997**, *16*, 64. [CrossRef] [PubMed]
36. Smith, D.P.; Rayter, S.I.; Niederlander, C.; Spicer, J.; Jones, C.M.; Ashworth, A. LIP1, a cytoplasmic protein functionally linked to the Peutz-Jeghers syndrome kinase LKB1. *Hum. Mol. Genet.* **2001**, *10*, 2869–2877. [CrossRef]
37. Walsh, M.F.; Ritter, D.I.; Kesserwan, C.; Sonkin, D.; Chakravarty, D.; Chao, E.; Ghosh, R.; Kemel, Y.; Wu, G.; Lee, K.; et al. Integrating somatic variant data and biomarkers for germline variant classification in cancer predisposition genes. *Hum. Mutat.* **2018**, *39*, 1542–1552. [CrossRef]
38. Spier, I.; Kerick, M.; Drichel, D.; Horpaopan, S.; Altmüller, J.; Laner, A.; Holzapfel, S.; Peters, S.; Adam, R.; Zhao, B.; et al. Exome sequencing identifies potential novel candidate genes in patients with unexplained colorectal adenomatous polyposis. *Fam. Cancer* **2016**, 281–288. [CrossRef]
39. Tripathi, M.K.; Deane, N.G.; Zhu, J.; An, H.; Mima, S.; Wang, X.; Padmanabhan, S.; Shi, Z.; Prodduturi, N.; Ciombor, K.K.; et al. Nuclear factor of activated T-cell activity is associated with metastatic capacity in colon cancer. *Cancer Res.* **2014**, *74*, 6947–6957. [CrossRef]
40. Wang, L.; Zhang, B.; Wolfinger, R.D.; Chen, X. An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genet.* **2008**, *4*, e1000115. [CrossRef]
41. Wang, L.; Chen, X.; Wolfinger, R.D.; Franklin, J.L.; Coffey, R.J.; Zhang, B. A unified mixed effects model for gene set analysis of time course microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **2009**, *8*, 47. [CrossRef] [PubMed]
42. Shirts, B.H.; Konnick, E.Q.; Upham, S.; Walsh, T.; Ranola, J.M.O.; Jacobson, A.L.; King, M.-C.; Pearlman, R.; Hampel, H.; Pritchard, C.C. Using Somatic Mutations from Tumors to Classify Variants in Mismatch Repair Genes. *Am. J. Hum. Genet.* **2018**, *103*, 19–29. [CrossRef] [PubMed]
43. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
44. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
45. Seshan, V.E.; Olshen, A. *DNAcopy: DNA Copy Number Data Analysis*, R package version 1.48.0. Bioconductor; Roswell Park Comprehensive Cancer Center: Buffalo, NY, USA, 2016.
46. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [CrossRef]
47. Guo, Y.; Li, J.; Li, C.-I.; Long, J.; Samuels, D.C.; Shyr, Y. The effect of strand bias in Illumina short-read sequencing data. *BMC Genom.* **2012**, *13*, 666. [CrossRef]
48. Chubb, D.; Broderick, P.; Dobbins, S.E.; Houlston, R.S. CanVar: A resource for sharing germline variation in cancer patients. *F1000Research* **2016**, *5*, 2813. [CrossRef]
49. Díaz-Gay, M.; Vila-Casadesús, M.; Franch-Expósito, S.; Hernández-Illán, E.; Lozano, J.J.; Castellví-Bel, S. Mutational Signatures in Cancer (MuSiCa): A web application to implement mutational signatures analysis in cancer samples. *BMC Bioinform.* **2018**, *19*, 224. [CrossRef]

50. Alexandrov, L.B.; Nik-Zainal, S.; Wedge, D.C.; Campbell, P.J.; Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **2013**, *3*, 246–259. [CrossRef]
51. Alexandrov, L.B.; Nik-Zainal, S.; Wedge, D.C.; Aparicio, S.A.J.R.; Behjati, S.; Biankin, A.V.; Bignell, G.R.; Bolli, N.; Borg, A.; Børresen-Dale, A.-L.; et al. Signatures of mutational processes in human cancer. *Nature* **2013**, *500*, 415–421. [CrossRef]
52. Forbes, S.A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C.G.; Ward, S.; Dawson, E.; Ponting, L.; et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **2017**, *45*, D777–D783. [CrossRef] [PubMed]

Supplementary Materials: Integrated Analysis of Germline and Tumor DNA Identifies New Candidate Genes Involved in Familial Colorectal Cancer

Marcos Díaz-Gay, Sebastià Franch-Expósito, Coral Arnau-Collell, Solip Park, Fran Supek, Jenifer Muñoz, Laia Bonjoch, Anna Gratacós-Mulleras, Paula A. Sánchez-Rojas, Clara Esteban-Jurado, Teresa Ocaña, Miriam Cuatrecasas, Maria Vila-Casadesús, Juan José Lozano, Genis Parra, Steve Laurie, Sergi Beltran, EPICOLON Consortium, Antoni Castells, Luis Bujanda, Joaquín Cubiella, Francesc Balaguer and Sergi Castellví-Bel

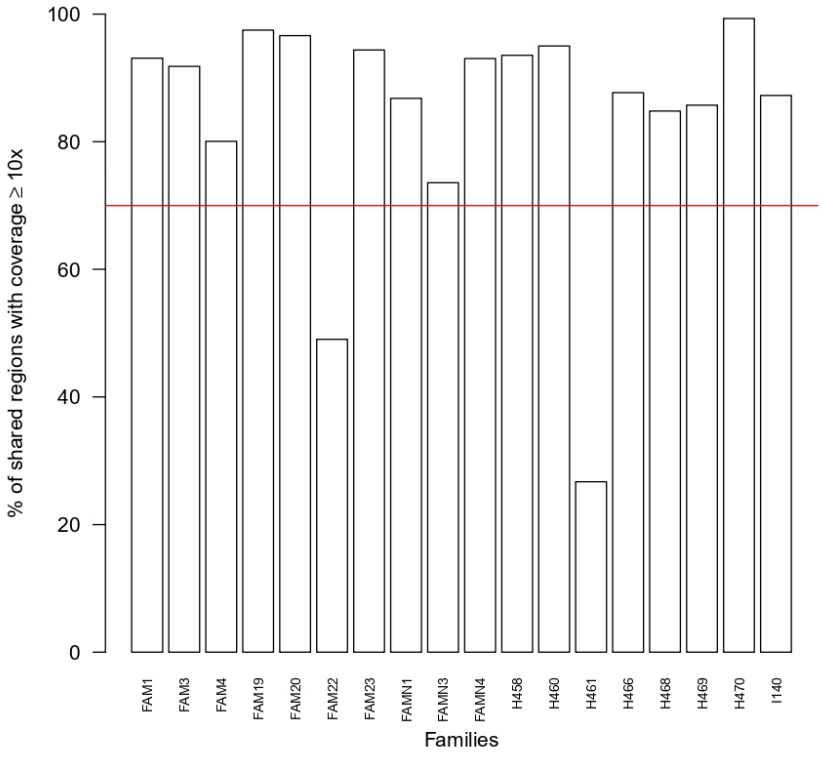
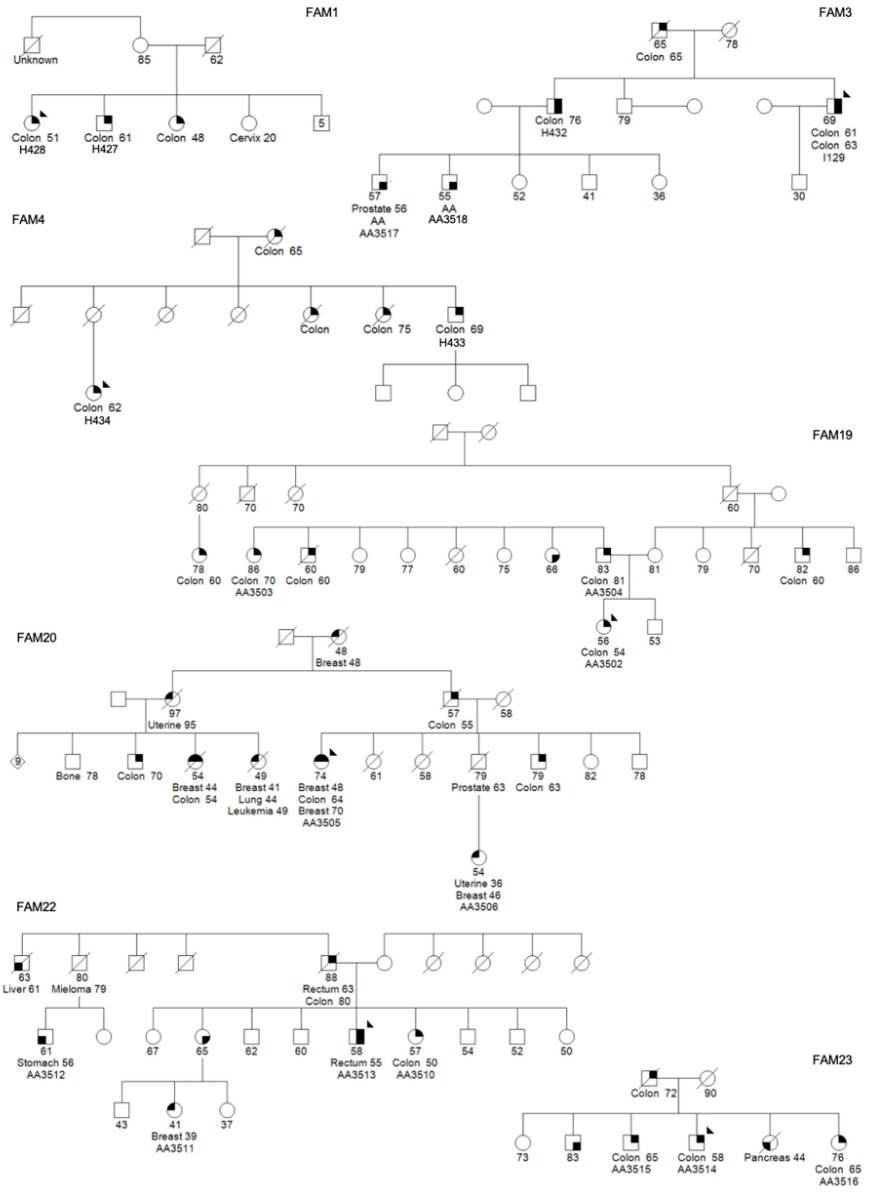
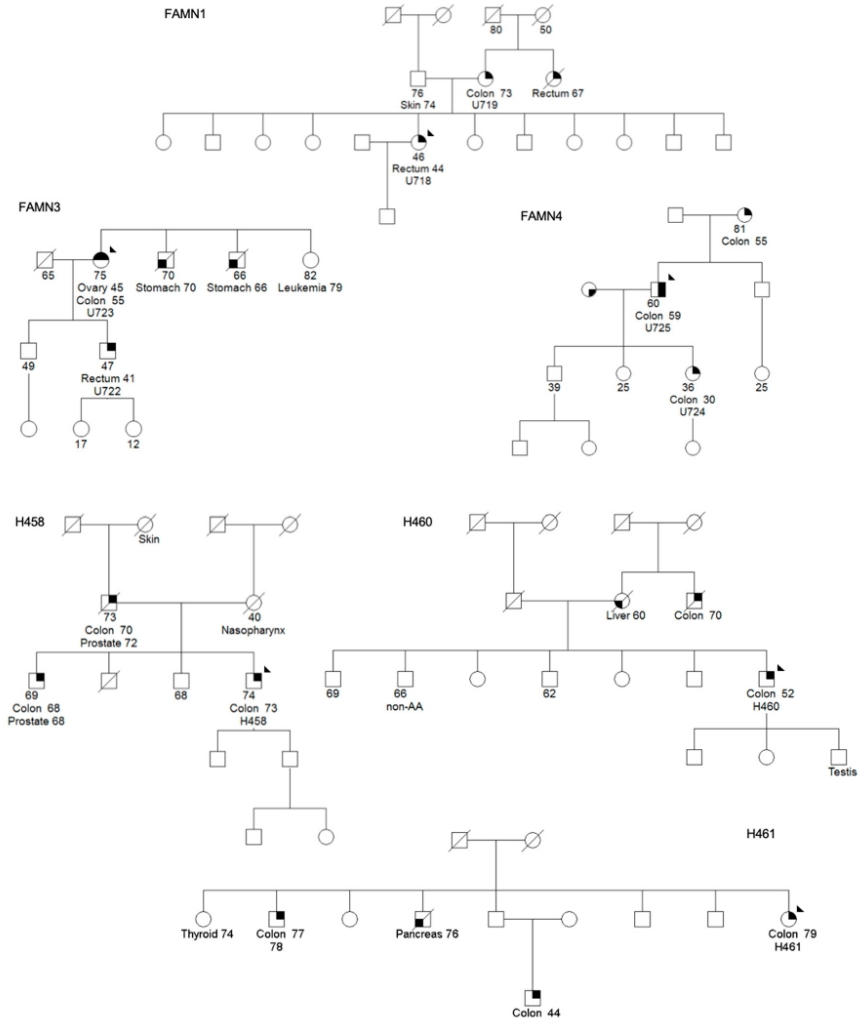


Figure S1. Histogram representing the percentage of genomic regions with a high-quality value of coverage (≥10×) with respect to all shared sequenced regions for each of the germline-tumor paired samples. Horizontal red line indicates sample filtering threshold (≥70% of shared regions with coverage above 10×).


FAM1
Unknown
85
62
Colon 51
H428
Colon 61
H427
Colon 48
Cervix 20
5
FAM3
65
Colon 65
78
Colon 76
H432
79
69
Colon 61
Colon 63
I129
57
Prostate 56
AA
AA3517
55
AA
AA3518
52
41
36
30
FAM4
Colon 65
Colon
Colon 75
Colon 69
H433
Colon 62
H434
FAM19
80
70
70
60
78
Colon 60
86
Colon 70
AA3503
60
Colon 60
79
77
60
75
66
83
Colon 81
AA3504
81
79
70
82
Colon 60
86
56
Colon 54
AA3502
53
FAM20
48
Breast 48
97
Uterine 95
57
Colon 55
58
9
Bone 78
Colon 70
54
Breast 44
Colon 54
49
Breast 41
Lung 44
Leukemia 49
74
Breast 48
Colon 64
Breast 70
AA3505
61
58
79
Prostate 63
79
Colon 63
82
78
54
Uterine 36
Breast 46
AA3506
FAM22
63
Liver 61
80
Mieloma 79
88
Rectum 63
Colon 80
61
Stomach 56
AA3512
67
65
62
60
58
Rectum 55
AA3513
57
Colon 50
AA3510
54
52
50
43
41
Breast 39
AA3511
37
FAM23
Colon 72
90
73
83
Colon 65
AA3515
Colon 58
AA3514
Pancreas 44
76
Colon 65
AA3516


FAMN1
80
50
76
Skin 74
Colon 73
U719
Rectum 67
46
Rectum 44
U718
FAMN3
65
75
Ovary 45
Colon 55
U723
70
Stomach 70
66
Stomach 66
82
Leukemia 79
49
47
Rectum 41
U722
17
12
FAMN4
81
Colon 55
60
Colon 59
U725
39
25
36
Colon 30
U724
25
H458
Skin
73
Colon 70
Prostate 72
40
Nasopharynx
69
Colon 68
Prostate 68
68
74
Colon 73
H458
H460
Liver 60
Colon 70
69
66
non-AA
62
Colon 52
H460
Testis
H461
Thyroid 74
Colon 77
78
Pancreas 76
Colon 79
H461
Colon 44

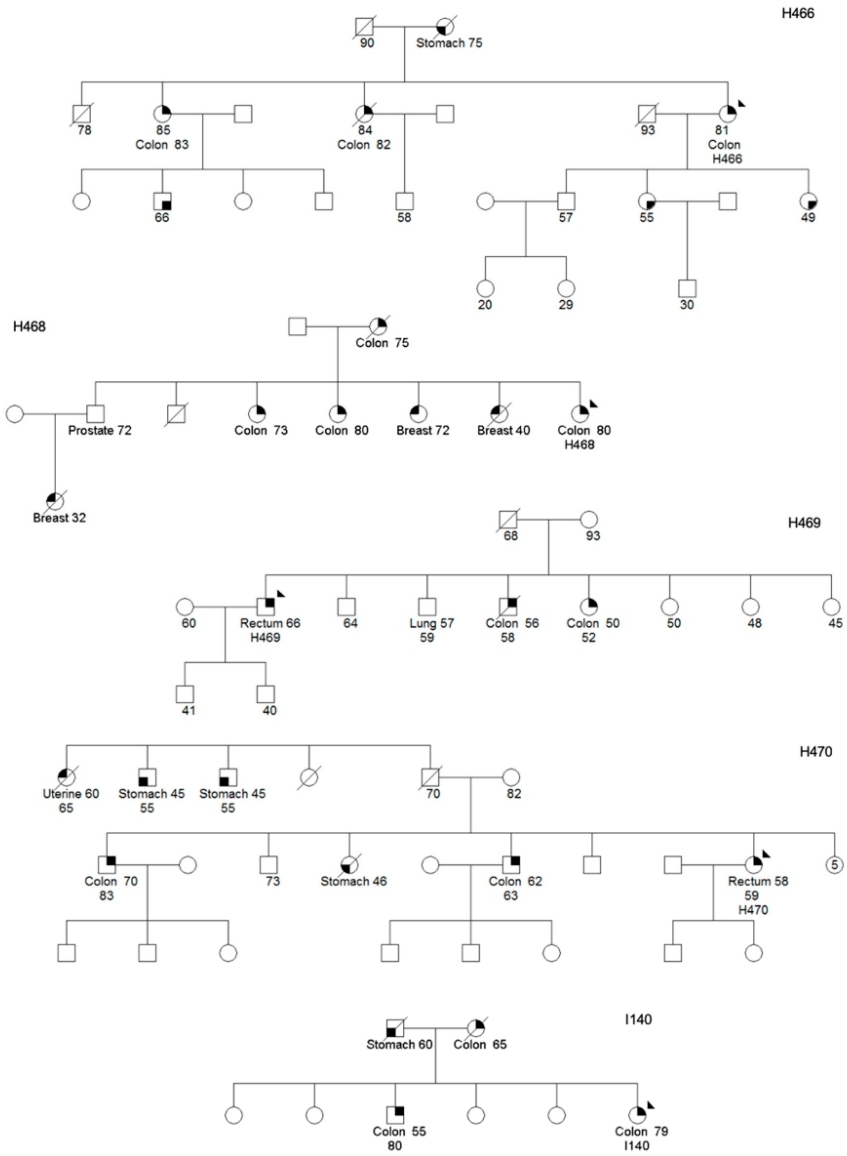


Figure S2. Pedigrees of the 18 families included in the study. Sample selected for germline and tumor whole-exome sequencing is indicated with an arrow. Filled symbols indicate affected for colorectal cancer (upper right quarter), adenoma/s (lower right quarter), gynecological cancer (ovary, uterine or breast cancer) (upper left quarter) and liver, stomach or pancreatic cancer (lower left quarter). Other cancer types are indicated in text with no symbol. IDs from samples undergoing germline whole-exome sequencing are also shown. AA/on-AA, advanced adenoma/non-advanced adenoma.

Table S1. Description of germline copy number variants detected after calling with CoNIFER and ExomeDepth.

Sample	CNV Class	Calling Tool	Location	Length (bp)	DGV Freq.	EPICOLON Freq.	Genes Included	Encoded Proteins Function/OMIM
AA3584	Deletion	CoNIFER	chr2:228,678,570-228,789,026	110,456	0	0	*CCL20, DAW1, SPHKAP*	Rheumatoid arthritis–*CCL20* (#601960); axonemal dynein –DAW1
		ExomeDepth	chr2:228,678,628-228,860,410	181,782	0	0		
AA3584	Duplication	CoNIFER	chr7:84,624,869-84,751,247	126,378	0	0	*SEMA3D*	Axon guidance (#609907)
		ExomeDepth	chr7:84,685,033-84,727,281	42,248	0	0		
AA3584	Duplication	CoNIFER	chr11:20,691,136-21,597,001	905,865	0	0	*NELL1*	Osteogenesis (#602319)
		ExomeDepth	chr11:21,555,920-21,596,568	40,648	0	0		
AA3589	Duplication	CoNIFER	chr5:140,529,839-140,555,957	26,118	0	1	*PCDHB6, PCDHB17, PCDHB7, PCDHB8*	Neural cadherins–*PCDHB7* and *PCDHB8* (#606333 and #606334)
		ExomeDepth	chr5:140,552,417-140,560,021	7604	9	0		
AA3589	Duplication	CoNIFER	chr9:117,085,336-117,088,757	3421	0	0	*ORM1*	Immunosuppression (#138600)
		ExomeDepth	chr9:117,085,943-117,087,432	1489	0	0		
AA3589	Deletion	CoNIFER	chr15:43,891,761-43,941,039	49,278	4	0	*CKMT1B, STRC, CATSPER2*	Mitochondrial creatine kinase–*CKMT1B* (#123290); deafness–*STRC* (#606440)
		ExomeDepth	chr15:43,888,606-43,897,597	8991	0	0		
U726	Duplication	CoNIFER	chr9:117,085,336-117,088,757	3421	0	0	*ORM1*	Immunosuppression (#138600)
		ExomeDepth	chr9:117,085,414-117,087,432	2018	0	0		

Abbreviations: bp, base pairs; CNV, copy number variant; DGV, Database of Genomic Variants; Freq., frequency; OMIM, Online Mendelian Inheritance in Man.

Table S2. List of genes where a potentially pathogenic germline SNV/indel and tumor LOH were identified in our samples.

Gene	Sample	Chromosome	Genomic Position	Reference Allele	Alternative Allele	Pathogenicity Tools	ExAC Freq.
ABCA12	AA3588	2	215,854,157	G	A	5/6	1.58E-05
ACADVL	AA3588	17	7,127,641	C	T	3/6	6.32E-05
AMOT	AA3588	X	112,022,894	C	T	3/6	6.87E-04
ANAPC5	AA3588	12	121,773,343	C	T	6/6	1.58E-05
ANO7	AA3588	2	242,144,358	C	A	NS	-
AREL1	AA3588	14	75,136,366	C	T	6/6	-
ARHGEF25	AA3584	12	58,008,801	G	C	SP	-
ARMC3	AA3584	10	23,321,876	T	C	3/6	2.53E-04
ASPM	AA3584, AA3582	1	197,073,396	A	C	5/6	1.11E-04
ASPN	AA3588	9	95,221,948	A	AT	FS	7.42E-04
ASTN1	AA3585	1	176,851,995	A	T	3/6	-
ATCAY	AA3585	19	3,907,776	G	A	4/6	3.99E-05
ATP4A	AA3588	19	36,046,582	G	A	4/6	1.97E-04
ATP7B	AA3583	13	52,515,271	C	T	3/6	5.58E-05
B4GALNT3	U729	12	670,538	G	A	6/6	8.68E-05
	AA3585	12	655,811	G	T	SP	5.74E-05
BCL9	AA3585	1	147,084,974	C	G	5/6	-
BLM	U729	15	91,306,382	C	T	6/6	7.89E-06
BORA	AA3585	13	73,319,166	C	A	6/6	2.96E-04
BRCA2	AA3599	13	32,913,453	GT	G	FS	-
CELSR1	AA3583	22	46,835,164	G	A	3/6	1.58E-05
CELSR2	AA3588	1	109,811,556	G	A	3/6	7.89E-05

CELSR3	AA3585	3	48,692,507	G	A	5/6	7.89E-06
	AA3589	3	48,688,833	C	T	4/6	1.11E-04
CLCA2	AA3588	1	86,913,377	C	T	6/6	1.82E-04
CNTNAP2	AA3588	7	147,259,253	G	A	4/6	-
COBLL1	AA3588	2	165,578,715	G	A	3/6	-
COL1A1	AA3586	17	48,270,048	C	T	3/6	1.74E-04
	AA3584	17	48,264,061	G	A	5/6	7.89E-06
COL6A3	AA3585	2	238,277,797	T	A	3/6	3.95E-05
COL9A3	AA3585	20	61,461,143	G	A	4/6	6.42E-05
COLEC12	AA3588	18	348,074	T	C	4/6	7.89E-06
CSMD3	AA3588	8	113,326,269	T	C	3/6	9.79E-04
CST6	AA3588	11	65,779,570	C	T	3/6	-
DCN	AA3588	12	91,545,375	AGGGTG	A	FS	7.10E-05
DHRS2	AA3585	14	24,108,386	A	G	SP	1.58E-05
DLG5	AA3584	10	79,580,910	C	T	3/6	1.58E-05
DNASE1	AA3589	16	3,706,649	G	A	6/6	4.82E-04
DOPEY1	U729	6	83,877,666	T	C	4/6	5.53E-05
DPT	AA3582	1	168,698,123	C	T	5/6	-
DST	AA3599	6	56,497,777	A	C	4/6	-
DYNC1H1	AA3588	14	102,500,421	C	A	6/6	1.50E-04
ECT2	AA3586	3	172,480,522	C	G	3/6	3.16E-05
ELK4	AA3582	1	205,589,296	G	C	4/6	9.47E-05
EPB41L4B	AA3596	9	112,017,891	A	C	5/6	-
EPPK1	AA3600	8	144,946,135	CTGAG	C	FS	8.05E-04

ERCC2	AA3584	19	45,867,712	C	T	4/6	-
FAT2	AA3600	5	150,946,850	A	G	5/6	1.58E-05
FDFT1	AA3582	8	11,679,329	G	A	6/6	3.16E-05
FERMT2	AA3582	14	53,417,220	G	A	3/6	3.08E-04
FGF12	AA3584	3	191,861,866	G	A	5/6	-
FMOD	AA3582	1	203,317,176	C	T	4/6	3.95E-05
FOXJ3	AA3582	1	42,657,223	C	T	3/6	4.34E-04
GNRHR	AA3585	4	68,619,623	G	A	6/6	7.89E-06
GPR110	AA3596, AA3582	6	46,976,836	TCA	T	FS	4.74E-05
GPR98	AA3582	5	90,059,209	G	A	3/6	8.81E-05
HCK	AA3598	20	30,671,808	C	T	6/6	7.89E-06
HEXIM1	AA3551	17	43,226,669	G	A	3/6	1.50E-04
HEY1	AA3588	8	80,679,521	C	A	3/6	-
HOOK3	AA3584	8	42,841,856	T	G	3/6	7.10E-05
HOXD4	AA3589	2	177,016,498	TC	T	FS	-
HSPG2	AA3551	1	22,202,391	C	T	3/6	3.16E-05
HUNK	AA3582	21	33,370,865	C	T	5/6	6.32E-05
IFT172	AA3585	2	27,672,571	C	T	6/6	1.34E-04
IGF2R	AA3589	6	160,412,298	G	A	6/6	7.89E-06
IL7R	AA3588	5	35,867,562	A	C	4/6	8.68E-05
ITGA4	AA3588	2	182,396,457	C	A	4/6	7.83E-04
ITIH5	AA3551	10	7,682,770	C	G	4/6	6.47E-04
KCNH5	AA3588	14	63,174,240	C	A	NS	1.06E-04
KNTC1	AA3589	12	123,068,956	T	C	6/6	1.28E-04

KRT23	AA3583	17	39,087,671	G	T	3/6	-
LATS2	AA3585	13	21,619,829	C	T	5/6	7.89E-06
LILRB2	U729	19	54,779,854	G	A	3/6	1.97E-04
LMO7	U726	13	76,335,093	C	G	5/6	-
LONP1	AA3583	19	5,705,801	T	C	3/6	4.74E-05
LTBP3	AA3598	11	65,315,172	C	A	4/6	-
MAML1	AA3582	5	179,193,441	C	A	4/6	1.74E-04
MAP3K4	AA3585	6	161,470,614	G	T	4/6	5.21E-04
MAST2	AA3551	1	46,496,378	G	T	3/6	2.55E-04
MCF2L	AA3588	13	113,744,418	C	T	3/6	-
MELK	AA3589	9	36,651,798	G	A	5/6	5.53E-05
MLLT4	AA3585	6	168,316,012	T	A	5/6	-
MYO1E	AA3584	15	59,510,112	C	T	6/6	2.37E-05
NID1	AA3585	1	236,145,007	C	T	4/6	7.89E-06
NISCH	AA3588	3	52,505,834	A	T	3/6	3.63E-04
NME7	AA3589	1	169,292,503	G	A	5/6	7.89E-06
NR1D1	AA3588	17	38,252,065	G	A	5/6	2.37E-05
NR3C2	U726	4	149,356,823	T	C	4/6	7.89E-06
NRCAM	AA3598	7	107,834,771	G	A	3/6	7.89E-06
NUP160	AA3598	11	47,840,937	G	C	5/6	1.50E-04
PARP2	AA3599	14	20,823,075	G	C	3/6	2.39E-05
PCDHB1	AA3584	5	140,433,067	A	T	5/6	7.89E-06
PCDHGA8	AA3596	5	140,773,936	C	T	4/6	1.58E-05
PDE1B	AA3584	12	54,968,980	T	G	6/6	-

PHGDH	AA3582	1	120,263,814	C	T	5/6	7.89E-06
PHKA2	AA3582	X	18,924,895	C	T	5/6	2.37E-05
PHRF1	AA3586	11	607,593	CGACT	C	FS	-
PI4K2A	AA3584	10	99,426,866	A	C	5/6	-
PIK3R3	AA3585	1	46,597,560	T	C	5/6	1.34E-04
PLCD3	AA3597	17	43,195,480	C	T	6/6	3.18E-05
PLEC	AA3584	8	144,998,705	C	T	3/6	1.85E-05
PLXND1	AA3598	3	129,308,229	C	G	5/6	1.11E-04
POSTN	AA3584	13	38,153,449	C	T	6/6	2.37E-05
PPFIA2	AA3582	12	81,734,962	C	T	4/6	8.00E-06
PREX2	U729	8	69,009,359	G	A	5/6	1.58E-05
PRRC2A	AA3584	6	31,592,082	C	A	3/6	9.55E-04
PSMD9	AA3583	12	122,337,659	A	T	3/6	2.37E-04
PSRC1	AA3551	1	109,823,551	G	A	4/6	6.31E-05
PTK6	AA3585	20	62,168,601	TC	T	FS	1.37E-04
PTPN14	AA3585	1	214,557,352	C	T	6/6	8.68E-05
PYGO1	AA3583	15	55,838,573	G	A	5/6	3.16E-05
RASSF6	AA3585	4	74,447,572	G	A	6/6	4.34E-04
RECQL	AA3589	12	21,643,302	C	A	6/6	-
	AA3589	12	21,643,306	G	T	4/6	-
RERGL	AA3589	12	18,234,381	A	G	6/6	4.58E-04
REV3L	AA3597	6	111,726,679	T	A	5/6	-
RIF1	AA3585	2	152,320,296	G	A	4/6	3.95E-05
RREB1	AA3586	6	7,231,360	G	C	5/6	7.12E-05

RRP12	AA3584	10	99,129,270	C	T	6/6	5.53E-05
SALL3	AA3582	18	76,757,135	C	T	5/6	-
SEC23B	AA3588	20	18,505,241	G	C	4/6	7.89E-06
SERPINB10	AA3585	18	61,587,044	C	T	3/6	6.79E-04
SHMT1	AA3585	17	18,232,669	C	T	6/6	4.74E-05
SIK3	AA3596	11	116,734,472	C	T	5/6	2.37E-05
SLC33A1	AA3582	3	155,571,417	G	A	5/6	2.68E-04
SLIT3	AA3582	5	168,199,939	G	C	6/6	-
SMARCA4	AA3597	19	11,096,021	C	T	5/6	7.89E-06
SORL1	AA3598	11	121,456,987	C	T	5/6	-
SPDL1	AA3584	5	169,021,633	A	G	5/6	4.58E-04
SPEG	AA3584	2	220,355,229	T	G	3/6	3.19E-04
SPTBN1	AA3585	2	54,895,645	G	T	3/6	-
STK11IP	AA3588	2	220,471,854	C	T	5/6	4.08E-04
TAF6	AA3582	7	99,708,920	T	A	3/6	-
TENC1	AA3586	12	53,453,362	C	T	3/6	5.53E-05
TENM2	AA3582	5	167,674,669	G	A	6/6	1.59E-05
TIAM2	AA3588	6	155,485,700	T	C	6/6	7.89E-06
TMBIM1	AA3600	2	219,140,257	A	G	6/6	1.74E-04
TMC6	AA3584	17	76,121,897	G	A	4/6	5.16E-05
TNRC6A	AA3600	16	24,800,913	A	G	4/6	8.68E-05
TRAP1	AA3598	16	3,708,209	G	A	3/6	7.89E-06
TRPM8	AA3551	2	234,854,632	C	T	6/6	3.95E-05
	AA3588	2	234,869,523	C	T	6/6	6.39E-04

VWA5A	AA3589	11	124,005,718	G	A	3/6	5.53E-05
YLPM1	AA3582	14	75,265,235	C	T	4/6	1.43E-04
ZFAND4	AA3584	10	46,122,195	A	T	5/6	8.45E-04
ZNF521	AA3582	18	22,804,982	C	T	4/6	3.95E-05

Abbreviations: ExAC, Exome Aggregation Consortium; Freq., frequency; FS, frameshift; NS, nonsense variant; SP, splicing-affecting variant.

Table S3. List of potentially pathogenic rare germline variants located in the final candidate genes in a cohort of 1006 familial early onset CRC patients from the CanVar database. Only frameshift or missense variants with CADD > 15 were selected. Significant variant enrichment in cases compared to ExAC control database (p-value < 0.05) have been marked in bold.

Gene	Genetic Variant	CADD	CanVar Freq.	ExAC Freq.	p-Value
ADCY8	p.Asp209Glu	23.4	0.00105	0.00002	**0.002**
	p.Gln320Glu	25.5	0.00060	-	**0.016**
	p.Ile675Val	25.5	0.00103	0.00002	**0.002**
	p.Val722Ile	18.83	0.00100	0.00008	**0.013**
	p.Ser812Leu	23.7	0.00050	0.00017	0.304
	p.Ala878Thr	16.76	0.00051	0.00007	0.138
	p.Asp893PhefsTer46	FS	0.00051	-	**0.016**
	p.Arg924Cys	33	0.00100	0.00002	**0.002**
BLM	p.Leu9Pro	26.2	0.00050	-	**0.016**
	p.Cys361Ter	FS	0.00050	-	**0.016**
	p.Glu880Gln	22.5	0.00050	0.00003	0.064
	p.Gly1359Glu	24.3	0.00061	-	**0.016**
BRCA2	p.Ile505Thr	18.18	0.00204	0.00071	0.060
	p.Leu1227GlnfsTer5	FS	0.00051	-	**0.016**
	p.Ser1230LeufsTer9	FS	0.00050	-	**0.016**
	p.Gly1529Arg	27.0	0.00050	0.00040	0.553
	p.Lys1690Asn	23.7	0.00050	0.00013	0.231
	p.Tyr1710Ter	FS	0.00050	-	**0.016**
	p.Leu2092ProfsTer7	FS	0.00050	0.00002	**0.048**
	p.Val2179AspfsTer10	FS	0.00050	0.00001	**0.032**
	p.Glu2856Ala	25.9	0.00550	0.00078	**<0.001**
	p.Thr3013Ile	22.1	0.00050	0.00022	0.358
	p.Tyr3035Cys	27.4	0.00053	0.00003	0.079
	p.Tyr3035Ser	26.9	0.00053	0.00006	0.123
	p.Leu3274Trp	28.7	0.00051	-	**0.016**
ERCC2	p.Arg143Gly	25.2	0.00053	0.00001	**0.032**
	p.Phe610LeufsTer99	FS	0.00052	0.00001	**0.032**
	p.Ala717Gly	24.8	0.00054	0.00033	0.490
HSPG2	p.Glu113Lys	24.2	0.00205	0.00018	**0.001**
	p.Gln372Arg	22.9	0.00109	0.00005	**0.007**
	p.Val376Ile	23.5	0.00052	0.00007	0.138
	p.Arg420Gln	32	0.00054	0.00025	0.409
	p.Asp746Val	20.7	0.00053	0.00001	**0.032**
	p.His779Tyr	23.3	0.00374	0.00094	**0.001**
	p.Asp802Tyr	29.0	0.00052	0.00001	**0.032**
	p.Ala803Thr	25.1	0.00052	0.00006	0.123
	p.Arg940Cys	32	0.00058	-	**0.016**
	p.Thr1182Met	25.1	0.00055	0.00004	0.094
	p.Glu1526Lys	29.2	0.00073	0.00002	0.064
	p.Val1736Ile	18.67	0.00053	0.00037	0.523
	p.Arg1758Gln	25.6	0.00052	0.00005	0.109

	p.Arg1779Trp	29.3	0.00054	0.00014	0.256
	p.Val1867Met	15.75	0.00058	-	**0.016**
	p.Ala1883Val	24.5	0.00114	0.00058	0.329
	p.Ala2164Val	22.5	0.00057	0.00002	**0.048**
	p.Gly2270Arg	29.7	0.00208	0.00026	**0.003**
	p.Arg2377His	25	0.00054	0.00009	0.179
	p.Ser2412Asn	16.42	0.00114	0.00072	0.658
	p.Leu2459His	25.5	0.00054	-	**0.016**
	p.Val2738Met	24.2	0.00050	0.00004	0.094
	p.Val3079Met	21.9	0.00160	0.00030	**0.027**
	p.Arg3159Gln	16.31	0.00057	0.00023	0.379
	p.Gln3188His	21.2	0.00267	0.00054	**0.006**
	p.Arg3334His	24.5	0.00059	0.00006	0.123
	p.Ala3396Val	15.38	0.00054	0.00024	0.389
	p.Leu3451Phe	28.8	0.00057	0.00040	0.553
	p.Pro3487His	26.7	0.00066	0.00020	0.337
	p.Gly3934Arg	26.2	0.00063	0.00002	**0.048**
	p.Arg4086Trp	33	0.00052	0.00061	1.000
PARP2	p.Val163Met	22.3	0.00050	0.00010	0.206
RECQL	p.Ile497AsnfsTer12	FS	0.00051	0.00008	0.152
	c.1667_1667+3delAGTA	SP	0.00051	0.00035	0.515
REV3L	p.Gln578Lys	22.2	0.00050	0.00001	**0.032**
	p.Trp1129Cys	32	0.00051	0.00002	**0.048**
	p.Asp1202Asn	22.7	0.00149	0.00034	**0.035**
	p.Cys1442Tyr	16.68	0.00050	-	**0.016**
RIF1	p.Arg116Cys	22.1	0.00298	0.00021	**<0.001**
	p.Ile224Val	15.61	0.00104	0.00043	0.220
	p.Pro403Leu	22.2	0.00050	0.00039	0.546
	p.Lys1303Asn	17.19	0.00249	0.00053	**0.005**
	p.Glu1598Ala	15.30	0.00050	0.00001	**0.032**
	p.Asp2378Glu	16.52	0.00050	-	**0.016**
SEC23B	p.Arg546Trp	26.1	0.00165	0.00005	**<0.001**
SMARCA4	p.Arg359Gln	16.81	0.00183	0.00020	**0.001**
STK11IP	p.Arg550Cys	24.3	0.00056	0.00001	**0.032**
	p.Arg1061Cys	28.1	0.00055	0.00002	**0.048**

Abbreviations: CADD, combined annotation dependent depletion; ExAC, Exome Aggregation Consortium; Freq., frequency.

DISCUSSION

Discussion

Both studies presented in this doctoral thesis involve the use of tumor mutational data, with the ultimate goal of the identification of new genes linked to germline predisposition to familial CRC. A user-friendly web application for somatic mutational profile characterization was developed and presented in the first study (Díaz-Gay et al., 2018). Conversely, an integrated germline-tumor WES data analysis was built and applied in a cohort of 18 familial CRC patients in the second study, along with a tumor mutational profile characterization using the previously developed application (Díaz-Gay et al., 2019).

Mutational Signatures in Cancer (MuSiCa) web application

The first published study presents the development of Mutational Signatures in Cancer (MuSiCa) application. MuSiCa is one of the first web tools available to perform a comprehensive somatic mutational profiling of human tumors sequenced with NGS techniques.

Calculation of TMB and reconstruction of somatic mutational profile are available in MuSiCa for samples provided by users. TMB is calculated as the number of somatic mutations per megabase sequenced, which is determined by the NGS approach used (WGS, WES or targeted sequencing). Regarding mutational signatures, version 2 signatures from COSMIC database (Wellcome Trust Sanger Institute, 2019a) were selected as the reference set for the refitting of the mutational profiles, since they were the consensus signatures used in most studies at the moment of the development of the application (Grolleman, Díaz-Gay, et al., 2019). It is worth to mention that COSMIC database was updated in May of 2019, with the arrival of a new set of consensus mutational signatures (version 3 (Wellcome Trust Sanger Institute, 2019b)). This update was possible thanks to recent worldwide collaboration projects which are sequencing large numbers of samples using NGS techniques. These novel signatures refer to SNVs similarly to v2 signatures, even though they also account for other variant classes, such as DBSs and indels (Alexandrov et al., 2019). A future update of MuSiCa would be needed in order to deal with this renewed state-of-the-art framework of mutational signature analysis.

Signature refitting analysis provided in MuSiCa is performed using the functionalities of the MutationalPatterns package. This R/Bioconductor library represented the most comprehensive software implementing mutational signatures at the time MuSiCa was conceived. Both *de novo* deciphering of new sets of signatures derived from the study samples and refitting analysis are allowed by MutationalPatterns (Blokzijl et al., 2018). This latter approach, implemented in MuSiCa, is based on solving

a NNLS optimization problem through the use of an active set method algorithm (Lawson & Hanson, 1974) included in pracma R package (Borchers, 2019).

MuSiCa provides a graphical user interface (GUI) to MutationalPatterns, specifically designed for non-specialized bioinformatic researchers, as well as some additional features. The user-friendly interface was developed by means of Shiny framework, designed to build interactive web applications directly from R code (W. Chang, Cheng, Allaire, Xie, & McPherson, 2019). MuSiCa is freely available as part of the website of our research group (http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html), which also benefits its straightforward use by any member of the scientific community. Additionally, since the application is hosted at the CIBERehd bioinformatic platform web server, no large computing resources are needed to perform the analysis, helping widespread dissemination. In fact, according to Google Analytics usage data retrieved for the first 14 months since the publication of MuSiCa article, 1,344 unique users from 53 different countries have accessed MuSiCa webpage, accounting for a total of 3,045 sessions (**Figure 20**). United States is the country with the largest number of users, around 30% of the total (407 out of 1,344 unique users), whereas Spain is second with approximately 10% (135 out of 1,344) of MuSiCa users in its first 14 months online. However, the number of researchers potentially benefiting from the developing of MuSiCa application for performing mutational signatures analysis could be even greater, considering that MuSiCa can also be executed locally. The instructions to be followed in this regard are freely available in the GitHub page of the project (https://github.com/marcos-diazg/musica). Required dependencies needed to install the application locally are presented in this website (as well as their specific versions), along with the source R code, which can be freely downloaded.

MuSiCa somatic mutational profiling is performed in a single sample basis, which provides great benefits in the case of small cohorts and individual samples (Blokzijl et al., 2018). Both scenarios are common in the clinical setting, where the mutational profile of every patient should be contrasted to the same set of consensus mutational signatures (Rosenthal et al., 2016; Baez-Ortega & Gori, 2019). Thus, MuSiCa is established as a useful tool for mutational signatures characterization in clinical practice, as long as NGS data from both germline and tumor DNA are available (in order to be able to identify somatic variants).

Other web applications addressing mutational signature analysis have been recently developed, also after MuSiCa publication. This is in agreement with the idea of spreading the benefits of this type of tumor characterization by improving the accessibility of the whole research community. Beneficiaries include basic and clinical scientists not particularly experts in bioinformatics (Baez-Ortega & Gori, 2019; Grolleman, Díaz-Gay, et al., 2019; Hanane et al., 2019).

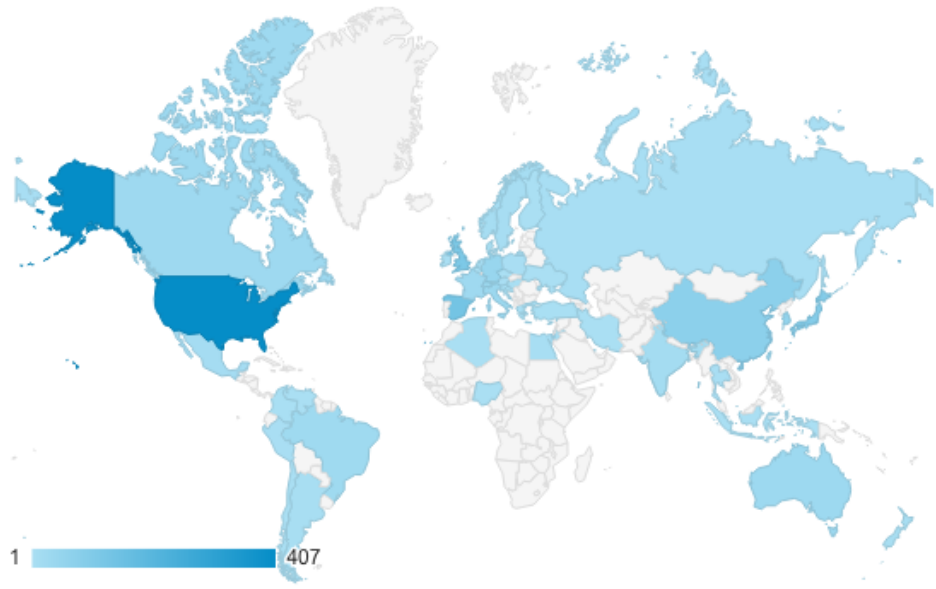


Figure 20. Worldwide distribution of MuSiCa users. Unique users accessing to the web version of MuSiCa application at http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html during the first 14 months from publication (June 14th 2018 – August 14th 2019).

First available tool displaying an online GUI was pmsignature. In this case, Shiny framework was also used for the development of the web application (https://friend1ws.shinyapps.io/pmsignature_shiny), which was presented at the same time that the software package. However, only *de novo* deciphering of mutational signatures was allowed by this application, according to its probabilistic model (Shiraishi et al., 2015).

MutaGene is a broad computational framework intended to provide a comprehensive characterization of tumor mutations and mutational associated process (https://www.ncbi.nlm.nih.gov/projects/mutagene). In this regard, the application enables users to analyze specific genes and look for potential driver mutations, as well as to explore mutational profiles and signatures and to perform signature refitting and sample comparison. MutaGene is basically intended for the examination of publicly available datasets, even though individual sample analysis is available through the *Identify* module. In this module signature refitting is performed using also NNLS as in the case of MuSiCa, even though only one sample at a time can be analyzed, thus not permitting the analysis of cohorts of more than one patient. This fact might be limiting the efficiency in some clinical settings when the comparison among a set of provided samples would be needed. Additionally, MutaGene allows the comparison of the uploaded sample against a full set of publicly available cancer samples, providing an estimation of the most probable cancer type and primary tumor site. Originally

conceived as a standalone website, MutaGene has recently been launched as a Python package in order to perform local analysis (Goncearenco et al., 2017).

DeconstructSigs, the other available R package dealing with signature refitting (Rosenthal et al., 2016), was also transposed to a freely accessible Shiny web application in the form of mSignatureDB (http://tardis.cgu.edu.tw/msignaturedb). Similar to MutaGene, this website provides a vast database of publicly available cancer samples, including 73 international projects (33 from TCGA and 40 from the International Cancer Genome Consortium) containing more than 15,000 samples. Apart from the exploration and comparison of these samples, mSignatureDB also permits the quantification of signatures at sample resolution (P.-J. Huang et al., 2018). As commented, deconstructSigs is the package selected to provide this refitting analysis, using a heuristic approach with ad hoc thresholds to solve the NNLS optimization. This implementation presents a much larger computation time than MutationalPatterns and, therefore, than MuSiCa, making it potentially impractical for a web environment (Blokzijl et al., 2018). In addition, *de novo* deciphering is also allowed by mSignatureDB by using the R implementation of NMF mutSignatures (Fantini et al., 2018), although this functionally does not seem particularly efficient for a web tool due to the large number of samples and computing time needed to perform an adequate and reliable analysis.

Mutalisk, published almost simultaneously to MuSiCa, is the most comprehensive web application regarding somatic mutational analysis at sample resolution to date (http://mutalisk.org). Apart from signatures decomposition (which can be performed either by linear regression or multinomial test), Mutalisk provides a panel of additional meaningful analyses, including localized hypermutation (also known as *kataegis*), transcriptional strand bias, GC content, DNA replication timing, histone modifications and DNase I hypersensitivity (J. Lee et al., 2018). Last three utilities make use of the Encyclopedia of DNA Elements (ENCODE) project information for data processing (Dunham et al., 2012). In addition, Mutalisk includes the option of using the new reference set of SNV signatures derived from latest studies (Alexandrov et al., 2019), which is currently hosted in COSMIC version 3 (Wellcome Trust Sanger Institute, 2019b).

With respect to input formats, MuSiCa is the most flexible application available. Apart from reference and mostly used Variant Calling Format (VCF), it allows the use of Tab-Separated Values (TSV), Excel and Mutation Annotation Format (MAF) files. This latter option is common to pack multi-sample data from TCGA projects hosted in the National Cancer Institute Genomic Data Commons. Conversely, Mutalisk only allows VCF format, whereas pmsignature uses a specific tab-separated format known as mutation position format, which can include one or more samples. Regarding MutaGene

and mSignatureDB, both work with VCF, MAF and TSV formats. Furthermore, MuSiCa presents some extra functionalities linked to sample classification when a set of samples is provided, which are not present in the other applications. Clustering and principal component analysis are included in order to find those samples with greatest similarities according to the signature reconstruction performed, which could have a great potential in certain clinical settings. An example of this impact could be achieved for example with a cohort of a certain cancer type, characterized by a very specific and well-defined phenotype, suspected to be caused by the same hereditary genetic defect. Comparison of mutational signature profiles of these patients with others of the same neoplasia could provide new insights about the predisposition mechanisms implicated in that particular subtype and, ultimately, guiding in the identification of the germline alteration responsible (Grolleman, Díaz-Gay, et al., 2019). This approach has recently been successfully used in the case of *NTHL1* deficiency and signature SBS30 association (Grolleman, de Voer, et al., 2019).

Some limitations are also shared among all the online applications currently available for mutational signature analysis. Regarding reference genomes, they are limited to human genomes GRCh37 and GRCh38. However, the analysis of other species widely used in biomedical research could be of great interest. In fact, mouse genomes are already supported by the Galaxy implementation of mutational signatures MutSpec (Ardin et al., 2016). On the other hand, refitting analysis is currently limited to SNV mutational signatures, even though new COSMIC reference signature framework also includes DBS and indel-associated signatures (Alexandrov et al., 2019; Wellcome Trust Sanger Institute, 2019b).

As a measure of the potential applicability of MuSiCa, somatic mutational profile characterization of colon tumors made as part of the TCGA project was replicated using mutational signatures (Muzny et al., 2012). A total of 433 samples were considered, while a multi-sample MAF file containing all the somatic variants was used as the input file, allowing a quick and straightforward uploading process into the application. Molecular profiling of somatic colon cancer was reproduced as in the original TCGA study, since key molecular pathways were deciphered in terms of their associated mutational signatures. Regarding MSI phenotype, it was linked with a cluster of samples presenting a predominance of a set of well-known signatures related to MMR deficiency (and the associated subsequent MSI acquisition), SBS6, SBS15, SBS20 and SBS26 (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Nagahashi et al., 2016; Drost et al., 2017; Alexandrov et al., 2019). A similar proportion of samples was found between those harboring MSI phenotype according to the original study and the cluster of samples mainly dominated by the MMR deficiency-associated signatures contribution in the mutational profile (Muzny et al., 2012). On the other hand, those samples harboring the highest TMBs were found forming a distinctive cluster primarily

dominated by signature SBS10. The etiology of this signature is the deficiency in polymerase epsilon (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013), which is in agreement with the TCGA study, since *POLE* alterations are a common feature of most ultrahypermutated CRC samples (Muzny et al., 2012). Remaining samples of the TCGA colon cancer cohort, representing most of the cases, have a mutational profile defined by the preponderance of age-associated signature SBS1 (Alexandrov et al., 2015). These samples correspond to those following CIN molecular pathway, where most notable genetic alterations are present in the form of somatic CNAs (Muzny et al., 2012; Carethers & Jung, 2015). As mutational profiles analyzed by MuSiCa only take into account SNVs, this fact results in a low value in TMB and most of the mutations being linked with *clock-like* mutational signatures (i.e. associated to the aging process). This is in accordance with the fact that CIN colon cancers are unequivocally associated with a low value of point mutations (Dienstmann et al., 2017). SNVs present in these samples would not presumably correspond to the driver alterations triggering the carcinogenesis development (this would be the case of the CNAs), but passenger events linked to the aging process and therefore with *clock-like* mutational signatures.

Integrated germline-tumor WES analysis of a familial CRC cohort

In the second study of this doctoral thesis, germline and tumor WES data from a cohort of 18 unrelated familial CRC patients was used in order to find novel candidate genes responsible for the germline predisposition to this neoplasia.

Samples used in this study represent a subset of a larger familial CRC cohort previously used by our research group and composed by 71 individuals from 38 families. On the basis of this cohort, some new candidates genes for predisposition to familial CRC were proposed in three recent studies considering different variant classes, including *BARD1*, *CDKN1B*, *EPHX1*, *NFKBIZ*, *SMARCA4* and *XRCC4* (Esteban-Jurado et al., 2015); *BRCA2*, *BRIP1*, *FANCC*, *FANCE* and *REV3L* (Esteban-Jurado et al., 2016); and *TMEM158* (Franch-Expósito et al., 2018). Selection criteria for the full cohort were mainly based on family history of CRC (modified Amsterdam II criteria) (Vasen et al., 1999). Specifically, families must have three or more relatives affected with CRC, two or more consecutive generations affected and at least one CRC case diagnosed before the age of 60 (Esteban-Jurado et al., 2015). This strategy of selecting families presenting strong aggregation for the disease has been commonly used in other studies looking for novel predisposition genes to CRC (DeRycke et al., 2013; Gylfe et al., 2013; Nieminen et al., 2014; Schulz et al., 2014; Seguí et al., 2015; Hansen et al., 2017), although cohorts consisting in unrelated early onset CRC cases have also been used (de Voer et al., 2013, 2016; Tanskanen et al., 2015; Brea-Fernandez et al., 2017). In addition, families were negative for germline mutations in common genes linked to well-known hereditary CRC syndromes, namely *APC* (associated to FAP) (Bodmer et al., 1987; Leppert et al., 1987),

MUTYH (MAP) (Al-Tassan et al., 2002) and the MMR genes *MLH1* (Lindblom et al., 1993), *MSH2* (Peltomaki et al., 1993), *MSH6* (Miyaki et al., 1997) and *PMS2* (Nicolaides et al., 1994) (Lynch syndrome). In addition, all tumors were MMR proficient (microsatellite stable). Germline WES data was available from the previous studies for the complete cohort of 71 patients. Additionally, in particular for this study, tumor DNA from formalin-fixed paraffin-embedded tissue was available for 18 of the patients (one per family), thus allowing matched germline-tumor WES in those samples. For these 18 families, apart from the matched sequenced patient, germline sequencing data from additional family members was available in some cases (three additional family members in 2 families, two in another 2 families and one in 6 families, whereas no extra data for the other 8 families).

Combined germline and tumor sequencing data allowed for the first time in our research group to test the profile of somatic genetic alterations, although with the main purpose of the identification of the germline defects responsible. In this regard, the experience accumulated in the analysis and identification of potential pathogenic variants of different classes in germline WES data, including SNVs, indels (Esteban-Jurado et al., 2016, 2015) and CNVs (Franch-Expósito et al., 2018), was exploited and translated to somatic sequencing data.

Thus, in-house pipelines developed in previous studies were used for the calling of all classes of germline alterations. SNVs and indels were initially identified using Genome Analysis Toolkit (GATK) HaplotypeCaller software, according to the well-known GATK Best Practices developed by the Broad Institute (integrated by both Massachusetts Institute of Technology and Harvard University) (DePristo et al., 2011; Van der Auwera et al., 2013). Conversely, CNVs were called using a combination of two tools, CoNIFER and ExomeDepth, both widely used in WES data according to different comparative studies (Guo et al., 2013; R. Tan et al., 2014; Kadalayil et al., 2015). In addition, a prioritization process was performed in all cases, in order to highlight those rare and potentially pathogenic alterations, affecting genes functionally compatible with a role in CRC susceptibility. In this regard, those variants not shared among all family members subjected to germline WES (when available) were also filtered out.

Likewise, with respect to somatic SNVs and indels, a similar strategy of prioritization looking for rare and potentially harmful variants was applied after the initial calling by MuTect2, the application from GATK also fulfilling GATK Best Practices. In addition, somatic LOHs were predicted by using the recently published ALFRED method, which allows to detect LOH regardless of which is the mechanism of origin, by using an allelic imbalance test (Park, Supek, & Lehner, 2018). LOH can be caused either by a CNV, in the form of a deletion, or by a uniparental disomy (UPD). LOH by UPD is found when a chromosome (or just a small part of it) containing the normal allele of a

heterozygous variant is lost and, subsequently, duplication of the remaining chromosome is produced. Accordingly, cell remains disomic but the mutated allele now is found in both chromosomes, i.e. the variant is present in homozygosis (Tuna, Knuutila, & Mills, 2009).

Integrated analysis of germline and tumor sequencing data allowed to test Knudson's two-hit hypothesis, looking for somatic alterations affecting genes already altered in germline DNA (Knudson, 1971). Thus, candidate TSGs must be harboring both a germline and a somatic genetic variation, leading to its complete loss of function. Those genes would be candidates to be implicated in germline predisposition to familial CRC, since their somatic inactivation would be triggering the neoplastic development. A similar strategy using Knudson hypothesis was used in some previous studies. Spier and collaborators analyzed a cohort of 7 patients with unexplained colorectal adenomatous polyposis, although no additional candidate gene was identified following the two-hit model. Nevertheless, only SNVs and indels were considered, therefore reducing the chances of finding the causative gene (Spier et al., 2016). On the other hand, in a recent analysis of more than 10,000 publicly available samples of different cancer types and considering exclusively LOH as the putative second hit, 13 genes were highlighted, including three well-known cancer predisposition genes, *BRCA1*, *BRCA2* and *ATM*, as well as new potential candidates for germline predisposition such as the histone methyltransferase *NSD1* (Park et al., 2018).

To the best of our knowledge, the described integrated analysis corresponds to the most comprehensive study of its kind to date, according to the different genetic variant classes taking into consideration. However, some limitations arise regarding other possible options to work as first germline or second somatic hits in the Knudson model. In this regard, lack of epigenetic data could have a great impact, since no information is available about different alterations such as aberrant methylations, histone modifications or non-coding RNAs, namely microRNAs or long non-coding RNAs (Khare & Verma, 2012; Okugawa, Grady, & Goel, 2015). Additionally, since the NGS technique used was WES, all non-coding regions of the genome were not assessed. This fact can have also an important effect on the discovery of potential alterations leading to CRC predisposition. Indeed, one of latest genes linked to inherited CRC, *GREM1*, was found implicated in hereditary mixed polyposis syndrome throughout a duplication located in the upstream promoter region of the actual gene that caused its overexpression (Jaeger et al., 2012). On the other hand, the prioritization step performed based on Knudson hypothesis testing by germline-tumor analysis is, in turn, limiting the identification of all possible TSG candidates for CRC predisposition in the analyzed samples. In this regard, for example genes harboring haploinsufficiency would not need a second mutational event in the tumor to trigger the cancer development (Deutschbauer et al., 2005), as it was proposed for candidate genes *BUB1* and *BUB3* (de

Voer et al., 2013). Furthermore, other strategies could be used in order to identify those genes more prone to be involved in the hereditary predisposition to familial CRC, including replication in additional cohorts or functional studies. These approaches have been successfully applied in some recent studies suggesting candidate genes such as *RPS20* (Nieminen et al., 2014), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *SETD6* (Martín-Morales et al., 2017) or *BRF1* (Bellido et al., 2018).

Prior to the deployment of the integrated germline-tumor WES analysis, quality control verifications were carried out for the sequencing data obtained from both DNA sources. In the case of germline samples, all were sequenced with good results, presenting a coverage over 95x in all cases. Conversely, two tumor samples were discarded for presenting low quality values. Shared exome regions sequenced with sufficient coverage among all available somatic samples were checked in this regard. A ratio lower than 70% was found in the two discarded samples.

Applying the corresponding germline pipeline for each of the different variant classes, 494 SNVs and 42 indels were found affecting germline DNA in the final cohort of 16 samples considered. In the case of germline CNVs, even if seven different alterations (five duplications and two deletions) were initially identified by the pipeline, no variant was finally taken into account, since the genes involved were not sufficiently functional compatible with predisposition to familial CRC (associated phenotypes and functions such as deafness, rheumatoid arthritis, immunosuppression or axon guidance were found among others). To complete the integrated analysis, somatic calling pipelines were applied to those genes harboring at least one of the identified germline alterations. A total of 143 genes were found to carry both germline and somatic variants (**Figure 21**). Three genes were identified to harbor an additional SNV in the somatic DNA, *ADCY8*, *HSPG2* and *TTN*. However, *TTN* gene was initially discarded for further analysis due to its vast length, which might be causing the accumulation of both germline and somatic variants simply by chance. In fact, *TTN* codifies for titin, the largest protein known. Titin is implicated basically in structural and mechanical functions, as well as in the regulation of cardiac and skeletal muscles (Chauveau, Rowell, & Ferreiro, 2014).

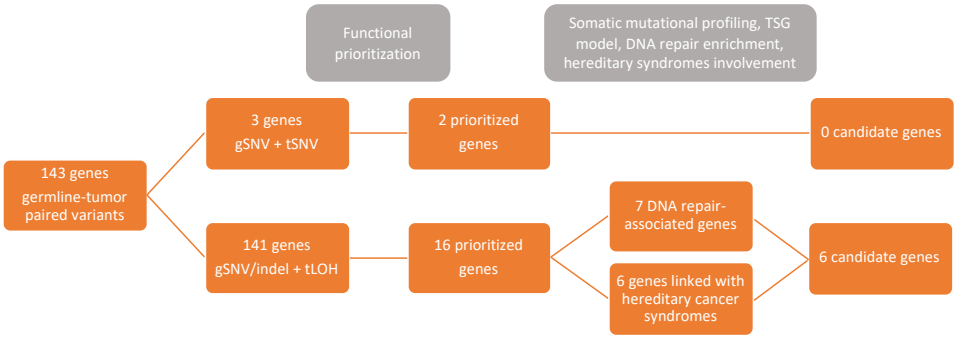


Figure 21. Schematic of gene prioritization process after the implementation of the integrated germline-tumor WES analysis. Different steps were performed to filter the initial list of genes harboring different genetic alterations both in germline and tumor DNA of a cohort of 16 familial CRC patients. Gene function (prioritizing DNA repair-associated genes), somatic mutational profiling, concordance with a tumor suppressor gene model of oncogenesis and involvement in hereditary cancer syndromes were considered to select the most suitable genes to be involved in germline predisposition to familial CRC. gSNV, germline single nucleotide variant; gSNV/indel, germline single nucleotide variant or small insertion or deletion; tLOH, tumor loss of heterozygosity; TSG, tumor suppressor gene.

On the other hand, 133 genes were predicted to carry a SNV in the germline genome followed by a LOH event in the tumor (including also the already commented *HSPG2*). In addition, eight genes shared a LOH as the second somatic hit, while an indel was identified as the first germline hit. Manual curation according to previously published functional knowledge was needed in order to reduce the number of genes to a first list of 16 potential candidates (**Figure 21**). Interestingly, an enrichment in DNA repair was found among the functions linked to the final selected genes, with seven out of 16 genes implicated. This is in agreement with part of classic predisposition genes to inherited CRC syndromes, such as *MUTYH* (part of BER pathway) (Al-Tassan et al., 2002) and *MLH1* (Lindblom et al., 1993), *MSH2* (Peltomaki et al., 1993), *MSH6* (Miyaki et al., 1997) and *PMS2* (Nicolaides et al., 1994) (MMR), as well as recent discoveries, as in the case of *POLE*, *POLD1* (polymerase proofreading) (Palles et al., 2013) and *NTHL1* (also BER) (Weren, Ligtenberg, et al., 2015). Along with DNA repair-associated genes, also those previously known to cause a cancer predisposition syndrome when mutated were highlighted, including *BLM* (associated with Bloom syndrome) (Ellis et al., 1995), *BRCA2* (hereditary breast and ovarian cancer syndrome) (Wooster et al., 1995), *ERCC2* (xeroderma pigmentosum) (Frederick, Amirkhan, Schultz, & Friedberg, 1994), *SMARCA4* (rhabdoid predisposition syndrome) (Schneppenheim et al., 2010). Interestingly, three of these four genes were also implicated in DNA repair processes (*BLM*, *BRCA2* and *ERCC2*) (AlDubayan et al., 2018). Additionally, two genes associated with well-known CRC predisposition syndromes were also considered after their identification in an ultrahypermutated sample, according to the somatic mutational profiling performed. On the one hand, *SEC23B* has been recently proposed as candidate

for Cowden syndrome, commonly caused by *PTEN* deficiency (Yehia et al., 2015). On the other hand, *STK11IP*, coding for an interacting protein to the well-known predisposition gene *STK11* (D. P. Smith et al., 2001) that is linked to Peutz-Jeghers syndrome (Giardiello et al., 1987). Interestingly, another known CRC predisposition gene, *PTEN* (Liaw et al., 1997), was also identified as a *STK11*-interacting protein (Mehenni et al., 2005). A total of 10 genes were prioritized by these two approaches (DNA repair and predisposition cancer syndromes) among all those harboring a combination of germline SNV/indel together with somatic LOH (**Figure 21**). DNA repair genes *PARP2* (implicated in BER pathway), *RECQL* (double-strand break repair via homologous recombination), *REV3L* (translesion DNA synthesis) and *RIF1* (double-strand break repair via nonhomologous end joining) were included besides the six already mentioned genes.

Thus, adding the two candidates carrying a different SNV in germline and tumor DNA, *ADCY8* and *HSPG2*, 12 genes were finally taken into consideration for final discussion. Subsequently, a case-control enrichment analysis was performed for these genes based on a cohort of familial early-onset CRC patients. Data from 1,006 patients stored in CanVar database was used (Chubb, Broderick, Dobbins, & Houlston, 2016), as well as normal controls for comparison from ExAC database (Lek et al., 2016). Rare and potential deleterious variants were found affecting all 12 genes, whereas an enrichment comparing to normal controls were detected for *ADCY8*, *BLM*, *BRCA2*, *ERCC2*, *REV3L*, *RIF1*, *SEC23*, *SMARCA4* and *STK11IP*.

Additionally, somatic mutational profiling was performed using the previously developed MuSiCa application (Díaz-Gay et al., 2018), in order to gain new insights for the prioritization of candidates for familial CRC predisposition. TMB, described as the number of SNVs per megabase sequenced, as well as mutational signatures contributions according to reference COSMIC signatures v2 (Wellcome Trust Sanger Institute, 2019a) were assessed in order to find putative relationships with the underlying germline alterations responsible. A total of 5 hypermutated tumors were found in the final cohort of 16 samples (showing more than 90 mutations per megabase), which is in agreement with the mentioned functional enrichment in DNA repair among the selected candidate TSGs, since DNA repair deficiencies are well-known major contributing factors in hypermutation (Campbell et al., 2017).

Regarding those genes selected for harboring germline-tumor SNVs (a different one in every analyzed genome), none of them was further consider as a putative candidate for inherited CRC. *ADCY8* is implicated in the generation of cyclic AMP from ATP, a pathway previously implicated in cancer although not following the TSG model expected in the integrated analysis based on Knudson's hypothesis. In this case, overexpression of one of the members of the gene family, *ADCY3*, was linked to the development of the neoplastic phenotype in gastric cancer cells (Hong et al., 2013). This

oncogenic role was also suggested for *HSPG2* in a study using CRC cell lines and tumor xenografts. *HSPG2*, encoding for a component of the extracellular matrix called perlecan, was identified as an inducer of tumor growth and angiogenesis (B. Sharma et al., 1998).

With respect to genes with a combination of germline SNV/indel and somatic LOH, six genes were finally prioritized among the 16 candidates, including well-known predisposition genes for additional neoplasias *BLM*, *BRCA2* and *ERCC2*, as well as DNA repair associated genes *RECQL*, *REV3L* and *RIF1*.

BLM and *RECQL* belong to the RecQ family of DNA helicases, responsible for double-stranded DNA unwinding and therefore with critical functions in DNA replication, recombination, transcription and repair (both in BER and double-strand break repair pathways) (Croteau, Popuri, Opresko, & Bohr, 2014). Their key role in cellular homeostasis is emphasized by the different cancer hereditary syndromes caused by biallelic germline deficiencies in three of the members of the family, *BLM*, Bloom syndrome (Ellis et al., 1995); *RECQL4*, Rothmund-Thompson syndrome (Kitao et al., 1999); and *WRN*, Werner syndrome (C.-E. Yu et al., 1996). In our cohort of 16 familial CRC patients, *BLM* was found carrying a germline missense variant (p.Pro690Leu) that was affecting its helicase domain and additionally predicted as pathogenic by all the *in silico* tools used. Somatic LOH was inferred as the somatic second hit leading to the putative inactivation of the protein. A low TMB was shown in the somatic mutational profiling of the affected patient, which could be indicating that the defect in the DNA repair is linked with a distinct carcinogenic mechanism different from the accumulation of somatic SNVs and indels. In fact, defects on *BLM* were previously associated to CIN and, therefore, with higher levels of somatic CNAs in mice (McDaniel et al., 2003; Chester, Babbe, Pinkas, Manning, & Leder, 2006). Additionally, *BLM* was proposed as candidate gene for predisposition to breast and CRC in some previous studies using WES for variant identification, although with a moderate-to-low penetrance (Thompson et al., 2012; de Voer et al., 2015). Thus, our study reinforces the potential role of *BLM* in hereditary CRC.

A complex germline variant was found in another patient in *RECQL* (p.Pro74_Trp75delinsGlnCys), composed by two single base substitutions separated by only three nucleotides, and therefore affecting two consecutive amino acids at protein level. As in the case of *BLM*, this variant was predicted as potentially deleterious according to the *in silico* tools assessed, whereas it was additionally not found in ExAC database. Apart from somatic LOH of *RECQL* gene, an hypermutated profile with near 100 mutations per megabase was found in the tumor of the affected patient. The hypothesis of the malfunctioning of a DNA repair mechanism resulting in a considerably high TMB was therefore proposed, even if *RECQL* deficiency was previously linked to

aneuploidy and CIN phenotype in mice (S. Sharma et al., 2007). *RECQL* has also been involved in double-strand break repair via non-homologous end joining (Parvathaneni, Stortchevoi, Sommers, Brosh, & Sharma, 2013), as well as in lengthening of telomeres without telomerase (Popuri et al., 2014). Interestingly, *RECQL* has been recently proposed as a novel candidate TSG for breast cancer predisposition (Cybulski et al., 2015), which strengthens the claim of this gene to be considered for familial CRC predisposition.

BRCA2 is a well-known predisposition gene, discovered to cause hereditary breast and ovarian cancer when mutations are present in the germline genome (Wooster et al., 1995). This gene, implicated in double-strand break repair via homologous recombination, has also been recently implicated in CRC susceptibility (Garre et al., 2015). In particular, the frameshift variant found in our cohort of familial CRC (p.Tyr1655fsTer15), also reported in one of the previous studies of the research group as part of the Fanconi anemia pathway (Esteban-Jurado et al., 2016), was classified as pathogenic in ClinVar database for hereditary breast and ovarian cancer syndrome. Additionally, up to four different breast cancer patients were found within the family carrying this mutation. Accordingly, this gene was selected as the one driving the predisposition to the different cancer types in this family, and was therefore communicated to the clinicians, in order to perform the appropriated genetic counseling. Consequently, this led to discard one of the other candidates, *PARP2*, since it was detected in the same family and *BRCA2* alteration was found in all sequenced affected relatives. This selection was also reinforced by the case-control analysis performed, since a significant enrichment in cases compared to controls was found for mutations in *BRCA2* but not in *PARP2*.

ERCC2 is the third candidate gene linked with a cancer predisposition syndrome. Double inactivation of this gene causes xeroderma pigmentosum, responsible for an increased susceptibility to skin cancer (Frederick et al., 1994). However, it has also been proposed as candidate for breast and ovarian cancer predisposition (Rump et al., 2016). *ERCC2* belongs to nucleotide excision repair (NER), particularly encoding for a DNA helicase in charge of double-stranded DNA unwinding near damage sites (Compe & Egly, 2012). A particular mutational signature, similar to COSMIC v2 mutational signature SBS5, has recently been associated to *ERCC2* deficiency in urothelial cancer (Kim et al., 2016). However, according to the mutational signature analysis performed using MuSiCa application, this signature was not present in the profile of the tumor carrying somatic *LOH* of *ERCC2*, as well as a predicted deleterious germline SNV (p.V230I). Conversely, predominance of mutational signature SBS1 was identified, as well as a subtle contribution of signature SBS7, linked to UV light exposure (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013; Hayward et al., 2017). Interestingly, bulky intrastrand DNA adducts formed by UV light are commonly repaired by NER pathway

(Marteijn, Lans, Vermeulen, & Hoeijmakers, 2014), thus leading to the hypothesis of the presence of this signature in this particular patient to be explained by the predicted *ERCC2* silencing.

REV3L encodes for the catalytic subunit of the DNA polymerase zeta and has an important role in translesion DNA synthesis. This polymerase is able to extend terminal mismatches in order to suppress stalled replication forks caused by DNA lesions (replacing common replicative polymerases delta and epsilon). Thus, DNA replication process is continued, although frequently incorporating genetic alterations. However, unrepaired DNA damage would be a potential source of replication and transcription errors, leading to double-strand breaks, CIN and ultimately cancer (Lange, Takata, & Wood, 2011; Yang et al., 2015; Zhao & Washington, 2017). Additionally, *REV3L* was associated with spontaneous tumor development in conditional knockout mice, therefore presenting a putative TSG role (Wittschieben et al., 2010). A potential deleterious germline missense variant, according to ExAC data and *in silico* prediction tools (p.Arg187Trp), was found in this gene, followed by the somatic inactivation via LOH of the wild type allele. Interestingly, a double inactivation of other candidate gene was also found in the same patient. It was the case of *SMARCA4*, involved in germline predisposition to rhabdoid tumors (Schneppenheim et al., 2010) and small cell carcinoma of the ovary, hypercalcemic type (Jelinic et al., 2014; Ramos et al., 2014; Witkowski et al., 2014). However, LOH validation by Sanger sequencing was available for this patient from previous studies of the research group (Esteban-Jurado et al., 2015, 2016). Somatic LOH was discarded for *SMARCA4* (Esteban-Jurado et al., 2015), whereas confirmed for *REV3L* (Esteban-Jurado et al., 2016), thus supporting this gene as a better candidate for familial CRC predisposition.

RIF1 is implicated in the non-homologous end joining DNA repair pathway. This pathway is involved in double-strand break repair, being the predominant repair mechanism for this type of DNA damage, although frequently introducing mutations during the process (H. H. Y. Chang, Pannunzio, Adachi, & Lieber, 2017). *RIF1* was suggested to contribute in the regulation of the pathway choice to repair DNA double-strand breaks (between non-homologous end joining and homologous recombination), via its interaction with *53BP1* (Escribano-Díaz et al., 2013). Again, a predicted pathogenic missense variant was found in the germline DNA (p.Arg1421His), followed by a predicted somatic LOH acting as second hit.

Somatic mutational profiling identified a particularly mutated tumor, with more than 500 mutations per megabase, thus leading to an ultrahypermutated phenotype (Campbell et al., 2017). Accordingly, a germline DNA repair defect was hypothesized to be the underlying cause of the inherited predisposition in that particular case. However, no gene belonging to this pathway was identified by the integrated germline-tumor

analysis. Interestingly, two genes were prioritized in this family, *SEC23B* and *STK11IP*, both carrying a combination of germline SNV and predicted tumor LOH, and proposed to be involved in well-known CRC predisposition syndromes, as previously commented (D. P. Smith et al., 2001; Yehia et al., 2015). However, a DNA repair germline deficiency was expected to be the underlying cause of the carcinogenic development in this sample, according to the ultrahigh TMB found in the somatic profiling. Therefore, these potential candidates were discarded for further discussion.

Regarding mutational signature analysis, a predominance of age-associated COSMIC v2 signature SBS1 was found in all samples, which is in agreement with the previous analysis of TCGA colon cancer samples made with MuSiCa. In the case of our cohort of familial CRC patients, as no alterations in MMR genes were present, a MMR proficient phenotype was assumed, commonly dominated by CIN phenotype and signature SBS1. As commented, somatic CNAs would have an important role as driver alterations in this type of tumors, which is in concordance with the somatic LOH testing performed in this study. However, this is in conflict with the high TMB values found across the whole cohort, with a median of almost 60 mutations per megabase and 5 out of 16 samples with more than 90 mutations per megabase (hypermutation is considered over 10 mutations per megabase (Campbell et al., 2017)). Conversely, CIN CRCs are mainly characterized by low TMB values (Dienstmann et al., 2017). Additional mutational signatures were also found in the familial CRC cohort, although with a much lower contribution. As previously mentioned, signature SBS7 caused by UV light exposure was identified, as well as signature SBS11. This latter signature is linked to alkylating agents, thus more related to treatment options, such as different chemotherapies, than predisposition. It is also important to mention that signature SBS11 is usually found to generate large numbers of somatic mutations (Wellcome Trust Sanger Institute, 2019b), which could be explaining the high TMB values found within the whole cohort. Interestingly, none of the known mutational signatures associated to DNA repair defects was found having a significant contribution to any sample, even if an enrichment in this cellular mechanism was identified among the prioritized genes by the integrated analysis.

The integrated germline-tumor WES analysis developed is in accordance with the recent recommendations from the Clinical Genome Resource through its recently established Germline/Somatic Variant Subcommittee. Both somatic mutational profiling and Knudson's two-hit hypothesis testing were considered in these guidelines, even if only TMB and mutational signatures analysis were suggested to be used in clinical settings on a routine basis. However, second hit assessment via LOH or second mutational event (SNV/indel) in the tumor were also recommended, although in a case-by-case basis and under the advisement of a multidisciplinary tumor-normal sequencing board in every cancer center (Walsh et al., 2018). Additionally, somatic mutational

profiling, particularly mutational signature analysis, was also recently suggested as a novel strategy to point out the most interesting candidate genes for cancer susceptibility also in the case of GWAS results (Chen et al., 2019).

Putative candidate genes for germline predisposition to familial CRC identified in this study could be helpful in future clinical practice, improving genetic counseling in affected families and benefitting early diagnosis. However, validation of the identified genetic alterations by orthogonal techniques, replication in independent familial CRC cohorts and further functional studies would be needed in order to confirm the association with CRC predisposition, as well as to provide new insights about the molecular mechanisms implicated.

CONCLUSIONS

Conclusions

Mutational Signatures in Cancer (MuSiCa) web application

1. Mutational Signatures in Cancer (MuSiCa) represents a user-friendly and freely available web application developed using the Shiny framework to perform somatic mutational profiling of a given set of cancer samples.

2. MuSiCa was established as one of the reference web applications for TMB calculation and mutational signatures characterization according to COSMIC reference signatures, being widely used since its publication.

3. Sample classification by clustering and principal component analysis according to the contributions of the different mutational signatures in a given set of provided samples is a unique feature of MuSiCa, not available in any other of the existing competitor applications to perform mutational signature analysis.

4. Molecular characterization of somatic CRC samples from TCGA project was accurately reproduced by mutational signature analysis with MuSiCa in a quick and user-friendly manner.

Integrated germline-tumor analysis of a familial CRC cohort

5. Integrated germline and tumor WES data analysis, considering different genetic variant classes and based on classic Knudson's two-hit hypothesis and somatic mutational profiling, was proved useful in the identification of new candidate TSGs involved in predisposition to familial CRC.

6. Six genes were identified as potential candidates for germline predisposition to familial CRC, including well-known predisposition genes in additional neoplasias *BLM*, *BRCA2* and *ERCC2*, as well as DNA repair-associated genes *RECQL*, *REV3L* and *RIF1*.

7. Somatic mutational profile analysis could be helpful to decipher the underlying responsible germline defect. In our study, it is exemplified by a candidate gene linked to DNA repair, *RECQL*, found mutated in the germline DNA of a sample harboring a hypermutated phenotype in the tumor, reinforcing the putative role of this gene in hereditary CRC.

SUMMARIES IN OTHER LANGUAGES

Identificación de novos xenes candidatos á predisposición xerminal a cancro colorrectal familiar mediante caracterización mutacional somática

Introdución

O cancro colorrectal (CCR) é unha das neoplasias malignas máis comúns e con maior mortalidade asociada no mundo, con máis dun millón e medio de novos casos e máis de 800.000 mortes cada ano (**Figura 1**) (Bray et al., 2018). A maior incidencia atópase nas rexións máis desenvolvidas, incluíndo Australia, Nova Zelandia, Europa, Asia Oriental e América do Norte (**Figura 2**). En Europa, o CCR representa o segundo tipo de cancro por incidencia e mortalidade considerando ambos sexos, mentres que en España é o primeiro en incidencia e só está detrás do cancro de pulmón en mortalidade (Ferlay et al., 2019). Como enfermidade complexa, a etioloxía da CCR implica a combinación de diferentes factores de risco. Ademais de factores non modificables, como a idade ou o xénero masculino, os factores ambientais foron tamén asociados cun aumento na incidencia de CCR, particularmente coa chamada occidentalización da dieta e do estilo de vida (Brenner et al., 2014).

O CCR foi un dos primeiros tumores sólidos caracterizados a nivel molecular, con diferentes vías de sinalización implicadas no inicio e na progresión da carcinoxénese (Fearon, 2011). Este proceso describiuse inicialmente a través da secuencia adenoma-carcinoma, onde unha acumulación de alteracións xenéticas en oncoxenes e xenes supresores de tumores (XSTs) dá como resultado unha transición dunha lesión precursora (chamada pólipo ou adenoma) a un carcinoma, a través de diferentes estados intermedios caracterizados por alteracións xenéticas e/ou epixenéticas específicas (**Figura 3**) (Vogelstein et al., 1988; Kuipers et al., 2015). Os oncoxenes defínense como aqueles xenes cuxa activación acelera o desenvolvemento do tumor, mentres que nos XSTs, pola contra, é a súa perda de expresión a que está ligada á adquisición do fenotipo neoplásico (Bashyam et al., 2019). Este fenotipo caracterízase principalmente por un crecemento incontrolado das células e a supresión dos mecanismos de morte e reparación celulares, así como pola adquisición das capacidades de invasión e metástase (**Figura 4**) (Hanahan & Weinberg, 2000, 2011). O defecto molecular inicial na maioría dos tumores colorrectais (máis do 70%) prodúcese no XST *APC*, provocando a desregulación da vía de sinalización Wnt/β-catenina (Kinzler & Vogelstein, 1996; Brenner et al., 2014), aínda que outras vías de sinalización tamén se ven afectadas durante a transformación neoplásica, incluíndo RAS-RAF-MAPK, PI3K-AKT, TGFβ e p53 (Kuipers et al., 2015). Recentemente, identificouse unha vía de

carcinoxénese colorrectal alternativa, iniciada por unha tipoloxía de lesións preneoplásicas diferenciadas, lesións serradas, que na actualidade se sabe que representan máis do 15% dos casos de CCR e que presentan características histolóxicas e moleculares diferenciadas con respecto aos adenomas convencionais (**Figura 3**) (Carballal et al., 2013; IJspeert et al., 2015).

A nivel molecular, considéranse tres vías principais para a carcinoxénese colorrectal: inestabilidade cromosómica (INC), inestabilidade de microsatélites (IMS) e a caracterizada por un fenotipo de hipermetilación de illas de dinucleótidos CpG (CIMP, polas súas siglas en inglés *CpG island methylator phenotype*) (**Figura 5**). A INC, caracterizada pola acumulación de alteracións no número de copia, foi a primeira vía molecular descrita e é coñecida por ser a orixe da maioría dos casos de CCR, especialmente dos casos esporádicos (ata un 85% destes últimos). En canto á IMS, esta defínese por alteracións nos microsatélites (secuencias repetitivas de ADN situadas ao longo do xenoma), que aparecen en forma de pequenas insercións ou delecións (indels), obtendo mutacións de terminación da proteína por cambio no patrón de lectura. Estas mutacións deberían ser corrixidas polo sistema de reparación do ADN denominado reparación de mal apareamento de bases (MMR, do inglés *mismatch repair*). Cando este sistema non funciona correctamente, aparece o fenotipo de IMS, amplamente utilizado como biomarcador para a detección dun MMR deficiente en CCR e ligado a hipermutación. Pola súa banda, o CIMP está ligado á hipermetilación de promotores de numerosos XSTs asociados ao cancro, o que provoca a supresión da súa transcrición (Carethers & Jung, 2015; Kuipers et al., 2015). Recentemente, describiuse unha nova clasificación molecular para o CCR baseada en patróns de expresión xénica, os chamados subtipos moleculares consenso (**Figura 6**) (Guinney et al., 2015; Dienstmann et al., 2017).

A predisposición xerminal a enfermidades complexas, como é o caso do CCR, implica unha distribución diversa de variantes xenéticas, que se poden clasificar segundo a súa frecuencia na poboación, así como respecto ao risco asociado a desenvolver unha certa enfermidade (coñecido como penetrancia) (**Figura 7**) (McCarthy et al., 2008; Manolio et al., 2009). As variantes de alta penetrancia defínense como as que causan un maior efecto na susceptibilidade á enfermidade, pero que normalmente son máis raras na poboación. Relaciónanse con enfermidades que seguen un patrón mendeliano (Mendel, 1866), onde a alteración dun só xene é frecuentemente responsable do fenotipo. Estas variantes foron identificadas clasicamente mediante estudos de ligamento, tamén no caso dos síndromes hereditarios de predisposición ao CCR (**Figura 8**) (Bodmer et al., 1987; Lindblom et al., 1993; Peltomaki et al., 1993). Por outra banda, as variantes de baixa penetrancia caracterízanse por ser comúns na poboación xeral e ter un pequeno efecto individualmente no desenvolvemento da enfermidade. Non obstante, unha combinación destas variantes, xunto coa interacción con factores de

risco ambiental pode contribuír significativamente á predisposición á enfermidade. Detectáronse principalmente por estudos de asociación do xenoma completo (GWAS, do inglés *genome wide association studies*), que no caso do CCR permitiron identificar ao redor de 130 variantes implicadas que explican un 7-8% da susceptibilidade asociada a esta enfermidade (**Figura 8**) (Jiao et al., 2014; Peters et al., 2015; Buniello et al., 2019). En determinadas enfermidades, como é o caso do CCR, a taxa de heredabilidade estimada con respecto a estudos clásicos en xemelgos e familias (12-35%) non está de acordo coa heredabilidade explicada polas variantes xenéticas cunha asociación coñecida coa enfermidade (2-8%), polo que isto implica unha heredabilidade *non filiada* (Jiao et al., 2014; Valle, Vilar, et al., 2019). Esta heredabilidade estaría relacionada en parte con aquelas variantes non o suficientemente frecuentes como para ser identificadas por GWAS, pero tampouco cun efecto sobre o desenvolvemento da enfermidade suficiente para seren detectadas por estudos familiares de ligamento (**Figuras 7-8**) (Manolio et al., 2009). Neste sentido, a secuenciación de nova xeración (SNX) desmarcouse como a ferramenta máis empregada para a identificación destas variantes. Esta técnica revolucionou o campo da xenética, permitindo identificar diferentes tipos de variantes implicadas na predisposición a diferentes enfermidades a un baixo custo relativo, incluíndo principalmente variantes dun só nucleótido (SNVs, do inglés *single nucleotide variants*) e indels, pero tamén variantes de número de copia (CNVs, do inglés *copy number variants*) (Lappalainen et al., 2019). As CNVs defínense como fragmentos de ADN dun tamaño superior a 50 nucleótidos con variacións no número de copias (delecións ou duplicacións) con respecto ao xenoma de referencia (Alkan et al., 2011). A aplicación máis exitosa da SNX en estudos biomédicos translacionais foi a secuenciación do exoma completo (SEC), é dicir, de todas as rexións codificantes do xenoma (Teer & Mullikin, 2010). Non obstante, para identificar novos xenes de predisposición, esta tecnoloxía require a implementación dunha estratexia de priorización, que permita reducir o elevado número de variantes que se identifican inicialmente (**Figura 9**) (Ott et al., 2015).

As síndromes hereditarias de predisposición ao CCR relacionadas con variantes xenéticas de alta penetrancia representan un 2-8% de todos os casos e ata un 6-10% se tamén se consideran as variantes de penetrancia moderada. Distintos xenes, pertencentes a diferentes vías de sinalización, foron implicados nestas síndromes, caracterizadas por seren orixinadas por distintas tipoloxías de lesións preneoplásicas (ou pólipos) (**Figura 10**) (Tomlinson, 2015). Clasifícanse fenotipicamente segundo a presenza ou non dunha acumulación destas lesións precursoras denominada polipose (**Figura 11**) (Valle, Vilar, et al., 2019). As síndromes polipósicas divídense á súa vez segundo o tipo de pólipos atopados nos pacientes. Con polipose adenomatosa están a polipose adenomatosa familiar e a súa variante atenuada (ligadas principalmente a mutacións xerminais no xene *APC*) (Leppert et al., 1987, 1990), a polipose asociada a

MUTYH (Al-Tassan et al., 2002), a polipose asociada á reparación do ADN por corrección realizada polas polimerases (ligada a defectos xerminais en *POLE* e *POLD1*) (Palles et al., 2013) e a síndrome tumoral asociada a *NTHL1* (implicada en ata 14 tipos tumorais diferentes) (Weren, Ligtenberg, et al., 2015). Por outra banda, provocada por pólipos serrados, aparece a síndrome da polipose serrada (da que só se propuxo un xene candidato para a súa predisposición hereditaria, *RNF43*, aínda que con controversia) (Gala et al., 2014); derivada de pólipos hamartomatosos, a síndrome de Peutz-Jeghers (ligada a defectos xerminais en *STK11*) (Giardiello et al., 1987), a síndrome de polipose xuvenil (*BMPR1A*, *SMAD4*) (Howe et al., 1998, 2001) e a síndrome tumoral *PTEN*-hamartoma / síndrome de Cowden (*PTEN*) (Liaw et al., 1997); e a través dunha combinación dos tres tipos de pólipos, a síndrome hereditaria de polipose mixta (*GREM1*) (Jaeger et al., 2012). Por outro lado, con respecto ás síndromes non polipósicas, destaca a síndrome de Lynch. Esta síndrome está asociada a mutacións xerminais nos xenes do sistema MMR (*MLH1*, *MSH2*, *MSH6*, *PMS2*) e constitúe a síndrome de CCR hereditario máis frecuente (H. T. Lynch et al., 2015). Debido a isto último, desenvolvéronse unha serie de guías clínicas para a identificación das familias con máis probabilidade de seren portadoras desta síndrome (**Figura 12**) (Vasen et al., 1999; Umar et al., 2004). As mencionadas síndromes presentan en xeral unha herdanza autosómica dominante, excepto no caso de aquelas ligadas a mutacións en xenes da vía de reparación do ADN por excisión de bases (BER, do inglés *base excision repair*), *MUTYH* e *NTHL1*, cuxo patrón de herdanza é autosómico recesivo (Valle, Vilar, et al., 2019).

Ademais das mencionadas síndromes de predisposición hereditarias (que explican ata un 8% da heredabilidade), especúlase con que os factores xenéticos estean detrás dun 12-35% do total de casos de CCR (Lichtenstein et al., 2000; Jiao et al., 2014; Peters et al., 2015). Esta heredabilidade *non filiada* foi amplamente estudada nos últimos anos co obxectivo de identificar novos xenes candidatos, que poderían ter un forte impacto no asesoramento xenético nas familias afectadas. A SNX foi a tecnoloxía empregada principalmente neste esforzo de identificación de novos xenes implicados na predisposición ao CCR (Valle, de Voer, et al., 2019). Así, un gran número de xenes candidatos foron propostos por diferentes grupos de investigación, incluíndo *BUB1*, *BUB3* (De Voer et al., 2013), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *BLM* (de Voer et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *MIA3* (Schubert et al., 2017), *SETD6* (Martín-Morales et al., 2017) e *BRF1* (Bellido et al., 2018) como os máis prometedores segundo os estudos funcionais realizados e a validación en cohortes de CCR familiar adicionais.

Segundo a hipótese dos dous *hits* de Knudson, a progresión neoplásica comeza con dous eventos mutacionais nun só xene (un XST), o que impide a súa expresión. Así, as diferenzas observadas entre as formas hereditarias e esporádicas/non hereditarias dun determinado cancro débense á diferente combinación destas alteracións xenéticas

ou *hits* (que poden ser de distintas clases: SNVs, indels, CNVs, perdas de heterocigosidade (LOH, do inglés *loss of heterozygosity*) ou alteracións na metilación). No caso dun cancro hereditario habería unha primeira alteración no ADN xerminal seguida dun segundo *hit* somático, mentres que nos casos esporádicos atoparíanse directamente dúas mutacións nas células tumorais (**Figura 13**). Deste xeito, explicaríase o frecuente diagnóstico a idades máis baixas dos cancros hereditarios, xa que só é necesario un evento mutacional no tumor para o desenvolvemento da enfermidade (Knudson, 1971).

Todos os cancros caracterízanse por múltiples mutacións somáticas. Así mesmo, estas mutacións clasifícanse en *driver* ou *passenger* segundo os seus efectos no desenvolvemento tumoral (Stratton et al., 2009). Aínda que na maioría dos estudos de secuenciación se priorizou a identificación de mutacións *driver*, por seren as seleccionadas positivamente e causantes da progresión carcinoxénica, as mutacións *passenger* tamén demostraron ser informativas. De feito, o número total de mutacións acumuladas por un tumor (denominado carga mutacional tumoral (TMB, do inglés *tumor mutational burden*)), moi variable entre tipos tumorais e tamén dentro do mesmo cancro (**Figura 14**) (Alexandrov, Nik-Zainal Wedge, Aparicio, et al., 2013), xurdiu nos últimos anos como un prometedor biomarcador para inmunoterapias, debido á súa relación coa carga de neoantíxenos (Chalmers et al., 2017).

Ademais da caracterización da TMB, as mutacións *passenger* tamén son as responsables da aparición dun novo campo de estudo nos últimos anos. Supoñendo que os patróns destas mutacións non varían co paso do tempo, pódense utilizar coma unha imaxe representativa dos mecanismos mutacionais que permaneceron activos durante o proceso carcinoxénico (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Cada proceso mutacional deixa unha marca particular no xenoma dunha célula, un perfil de mutacións específico denominado sinatura mutacional. Mecanismos celulares endóxenos, como a replicación e a reparación do ADN, poden xerar mutacións debido á súa taxa intrínseca de erro. Doutra banda, as mutacións tamén poden deberse a exposicións mutaxénicas exóxenas, como sería o caso do tabaco ou da luz ultravioleta. Así, o conxunto final de mutacións recollidas nun tumor está determinado pola intensidade e duración de todos os procesos mutacionais activos durante a progresión neoplásica (**Figura 15**) (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

O dano ao ADN pode aparecer baixo diferentes tipos de variantes xenéticas, aínda que para a descrición das sinaturas mutacionais ata o momento usáronse principalmente SNVs por motivos técnicos. Así, no conxunto actual de sinaturas mutacionais de referencia considéranse seis tipos de cambio de nucleótido, segundo a pirimidina mutada da parella de bases de Watson-Crick, incluíndo catro posibles

transversións, C>A, C>G, T>A e T>G, e dúas transicións, C>T e T>C. Para unha caracterización máis estrita dos procesos mutacionais responsables das mutacións, tamén se teñen en conta as bases adxacentes ao cambio nos contextos 5' e 3', dando lugar a un total de 96 posibilidades (6 substitucións de bases * 4 nucleótidos anteriores * 4 nucleótidos posteriores) (**Figura 16**). Deste xeito, cada sinatura mutacional está composta por unha distribución única destes 96 posibles tipos de mutacións (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). Nos últimos anos xerouse un modelo matemático que permitiu a detección e cuantificación precisa de cada unha das sinaturas mutacionais asociadas aos diferentes procesos mutaxénicos implicados no cancro. Para isto, utilizouse inicialmente un algoritmo baseado na factorización matricial non negativa chamado SigProfiler, que foi implementado usando MATLAB (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Recentemente, este algoritmo foi traducido a outras linguaxes de programación abertas (Gehring et al., 2015; Blokzijl et al., 2018), mentres que tamén apareceron novas estratexias computacionais para a identificación de sinaturas mutacionais (Kasar et al., 2015; Shiraishi et al., 2015; Baez-Ortega & Gori, 2019).

A través destes modelos computacionais foi posible a extracción de sinaturas mutacionais de referencia, cada unha delas asociada a un proceso mutaxénico específico a partir do cal, nalgúns casos, se identificou a súa etioloxía. Estas sinaturas de referencia permiten que a análise de sinaturas mutacionais non só se restrinxa á identificación agnóstica de novas sinaturas (coñecida como análise *de novo*, é dicir, sen utilizar ningún coñecemento previo). Tamén fan posible a caracterización dos procesos mutacionais implicados a nivel de mostra con respecto a unha referencia (denominada análise de axuste de sinaturas mutacionais). Para levar a cabo esta tipoloxía de análise, máis orientada á súa aplicación na práctica clínica, tamén apareceron novas ferramentas bioinformáticas nos últimos anos (Rosenthal et al., 2016; Blokzijl et al., 2018). Non obstante, aínda están orientadas a expertos en bioinformática, sendo inaccesibles a unha parte importante da comunidade científica. O número de sinaturas de referencia foi crecendo paulatinamente, a medida que aumentou o número de mostras de tumor analizadas, o que se debe ao aumento sucesivo da potencia estatística do modelo matemático. Nunha primeira aplicación desta metodoloxía extraéronse cinco sinaturas mutacionais de SNVs (posteriormente reducidas a catro despois dunha optimización do modelo) dunha cohorte de 21 mostras de cancro de mama (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Campbell, et al. , 2013). A continuación, o conxunto de referencia de sinaturas mutacionais ampliouse a 21 (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013), e máis tarde a 30 despois da aplicación desta metodoloxía a unhas 12.000 mostras de 40 tipoloxías diferentes de cancro (**Figura 17**) (Alexandrov et al., 2015; Tate et al., 2018). Este conxunto de 30 sinaturas mutacionais de referencia foi empregado na gran maioría das publicacións que realizaron análise de sinaturas

mutacionais nos últimos anos (Grolleman, Díaz-Gay, et al., 2019) e pódese atopar como parte da base de datos COSMIC (versión 2 – marzo 2015) (Wellcome Trust Sanger Institute, 2019a). Finalmente, o conxunto actual de sinaturas mutacionais de referencia foi extraído de máis de 23.000 mostras de cancro e consta de 49 sinaturas de SNVs (tamén chamadas sinaturas SBSs, do inglés *single base substitutions*), mentres que incorporou tamén outras tipoloxías de variantes, incluíndo 17 sinaturas asociadas a indels e 11 ligadas a substitucións de dúas bases consecutivas (**Figura 18**) (Alexandrov et al., 2019). Tamén está dispoñible en COSMIC (versión 3 – maio de 2019) (Wellcome Trust Sanger Institute, 2019b).

Atopouse unha distribución diferente de sinaturas mutacionais nos diferentes tecidos, o que está de acordo coas diferentes taxas de substitución celular, así como coa distinta influencia das exposicións ambientais segundo o tecido en cuestión. Algunhas sinaturas, como é o caso de SBS1, SBS5 e SBS40, asociáronse coa idade de diagnóstico, polo que reflicten a influencia do proceso de envellecemento na carcinoxénese (Alexandrov et al., 2015, 2019). En canto ao CCR, segundo a información dispoñible en COSMIC e unha serie de publicacións recentes, hai unha contribución de diferentes sinaturas mutacionais, incluídas as mencionadas sinaturas relacionadas co envellecemento. Non obstante, estas sinaturas contribúen a un número reducido de mutacións, en comparación coas relacionadas especificamente con dous coñecidos defectos moleculares presentes no CCR: as deficiencias nos procesos de reparación do ADN por MMR e por corrección das polimerases (7 sinaturas diferenciadas no caso dun mal funcionamento do sistema de MMR: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 e SBS44, e as sinaturas SBS10a e SBS10b no caso de mutacións no dominio exonuclease da polimerase *POLE*) (Nagahashi et al., 2016; Alexandrov et al., 2019). Adicionalmente, algunhas destas sinaturas asociáronse coa concorrencia de alteracións xenéticas nos dous sistemas de reparación do ADN, incluíndo as sinaturas SBS14 (mutación en *POLE* e nos xenes do MMR) e SBS20 (mutación en *POLD1* e MMR defectuoso), atopándose tamén esta última en casos de CCR (Haradhvala et al., 2018; Alexandrov et al., 2019). O espectro de sinaturas mutacionais asociadas ao CCR foi ampliado recentemente coa inclusión de dúas sinaturas relacionadas co sistema BER de reparación do ADN e especificamente con defectos en dous xenes de predisposición ao CCR coñecidos: *MUTYH* (sinatura SBS36) (Pilati et al. ., 2017; Viel et al., 2017) e *NTHL1* (SBS30) (Drost et al., 2017; Grolleman, de Voer, et al., 2019). Outras sinaturas tamén foron asociadas ao CCR, aínda que cun papel menos predominante, como é o caso de SBS2, SBS13 (ligadas á actividade das deaminases APOBEC), SBS3 (sistema de reparación do ADN por recombinación homóloga defectuoso e mutacións en *BRCA1/2*), SBS9 (actividade da polimerase eta), SBS17a, SBS17b, SBS18 (danos no ADN causados por especies reactivas do osíxeno), SBS12, SBS28, SBS37 e SBS41 (etioloxía descoñecida), así como

unha nova sinatura identificada por Roerink e colaboradores nun estudo recente (Nagahashi et al., 2016; Roerink et al., 2018; Alexandrov et al., 2019).

As sinaturas mutacionais, así como o TMB, pódense utilizar para a identificación de defectos xenéticos xerminais que estiveron activos durante a orixe e a evolución dun certo cancro. Isto é particularmente evidente para aquelas sinaturas mutacionais cunha etioloxía coñecida e, en particular, para as asociadas a procesos mutacionais responsables de síndromes hereditarias de predisposición ao cancro, coma os derivados de defectos nos mecanismos de reparación do ADN (**Figura 19**) (J Ma et al., 2018; Van Hoeck et al., 2019). No caso dos sistemas de corrección das polimerases e MMR, os defectos xerminais identificáronse ligados a unha TMB elevada, é dicir, a tumores hipermutados, e ademais ás mencionadas sinaturas mutacionais características (Muzny et al., 2012; Kandoth et al., 2013; Alexandrov et al., 2019). As sinaturas mutacionais, así como a análise da TMB, poderían axudar na identificación e descubrimento dos procesos mutacionais responsables das distintas síndromes hereditarias de cancro, así como favorecer o diagnóstico xenético e a selección de tratamentos nos pacientes, como se demostrou recentemente no caso das alteracións nos xenes *BRCA1/2* e *NTHL1* (Davies et al., 2017; Grolleman, de Voer, et al., 2019).

Hipótese

O CCR é unha enfermidade complexa e, polo tanto, cunha etioloxía na que se entrelazan factores xenéticos e ambientais. A predisposición xenética está detrás de ata un 35% dos CCRs segundo estudos familiares e de xemelgos, mentres que as síndromes de predisposición coñecidas e asociadas a defectos xenéticos xerminais específicos só explican un 2-8% dos casos. Deste xeito, obsérvase unha heredabilidade *non filiada* para esta neoplasia. A SNX é a técnica máis adecuada para levar a cabo a identificación de novos xenes implicados na predisposición ao CCR, como se demostrou en estudos recentes en xenes como *POLD1*, *POLE* e *NTHL1*. Non obstante, esta tecnoloxía identifica un gran número de variantes xenéticas en cada paciente, xerando así a necesidade dunha estratexia de priorización. Neste sentido, segundo a hipótese clásica dos dous *hits* de Knudson, ademais das alteracións xenéticas xerminais, tamén as somáticas poden desempeñar un papel fundamental para achegar novos coñecementos respecto á predisposición hereditaria ao CCR. Así, a análise do perfil mutacional somático foi utilizada recentemente para a identificación de novos xenes de predisposición a CCR, así como un biomarcador prometedor para o diagnóstico, prognóstico e tratamento desta neoplasia. Aínda que se desenvolveron diversos paquetes bioinformáticos para realizar este tipo de análises, segue sendo inaccesible para unha proporción substancial da comunidade científica.

Obxectivos

O obxectivo principal da presente tese de doutoramento é identificar novos xenes candidatos que poidan estar implicados na predisposición xerminal ao CCR familiar. Co fin de seren utilizadas como estratexias de priorización, desenvolveranse tanto unha análise combinada xerminal-tumoral de datos de SEC como unha aplicación bioinformática para realizar caracterización mutacional somática.

Con este fin, realizaranse os seguintes obxectivos específicos:

1. Desenvolvemento dunha aplicación informática para a análise de perfís mutacionais somáticos, a través dunha interface sinxela axeitada para investigadores non especializados en bioinformática e de libre acceso a través dunha páxina web. Tanto a caracterización da TMB como o axuste das sinaturas mutacionais segundo as sinaturas de referencia versión 2 de COSMIC estarán dispoñibles, así como a clasificación de mostras mediante *clustering* e análise de compoñentes principais.

2. Análise integrada, baseada na hipótese dos dous *hits* de Knudson, de datos de SEC procedentes de ADN xerminal e tumoral dunha cohorte de 18 pacientes de CCR familiar, co obxectivo de identificar novos XSTs potenciais. Teranse en conta distintas clases de alteracións xenéticas, mentres que os xenes candidatos serán seleccionados cando tanto o ADN xerminal como o tumoral estean afectados por unha destas alteracións.

3. Caracterización somática mutacional da mencionada cohorte de CCR familiar mediante a ferramenta bioinformática xerada anteriormente, a través da análise da carga mutacional tumoral e as sinaturas mutacionais.

Resultados e discusión

O primeiro dos estudos publicados como parte desta tese de doutoramento presenta o desenvolvemento da aplicación MuSiCa (do inglés *Mutational Signatures in Cancer*), que constitúe unha das primeiras ferramentas web dispoñibles para realizar unha caracterización mutacional somática completa dos tumores secuenciados con técnicas de SNX.

Tanto o cálculo da TMB como a reconstrución dos perfís mutacionais somáticos segundo as sinaturas mutacionais de referencia versión 2 de COSMIC (Wellcome Trust Sanger Institute, 2019a) están dispoñibles en MuSiCa. Será necesaria unha futura actualización da aplicación para a adaptación á nova versión 3 destas sinaturas de referencia, que debería incluír as novas clases de variantes a ter en conta. En canto ao axuste de sinaturas mutacionais, MuSiCa utiliza como base o paquete de R/Bioconductor MutationalPatterns (Blokzijl et al., 2018), que está baseado na resolución dun problema de optimización de mínimos cadrados non negativos mediante un algoritmo de método de conxunto activo (Lawson & Hanson, 1974) incluído no

paquete de R pracma (Borchers, 2019). MuSiCa proporciona unha interface gráfica a este paquete, creada a través do paquete de R Shiny (W. Chang et al., 2019) e deseñada especificamente para investigadores non especializados en bioinformática, así como algunhas características adicionais. MuSiCa está dispoñible de xeito gratuíto como parte da páxina web do noso grupo de investigación (http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html), o que permite o seu uso de forma sinxela por parte de calquera membro da comunidade científica, sen necesidade de grandes recursos a nivel informático. De feito, segundo os datos recollidos pola plataforma Google Analytics durante os primeiros 14 meses desde a publicación de MuSiCa, 1.344 usuarios únicos de 53 países diferentes accederon ao sitio web da aplicación, facendo un total de 3.045 sesións (**Figura 20**). Tamén é posible utilizar MuSiCa de forma local, para o cal as dependencias requiridas para a súa instalación, así como o código fonte en R, están dispoñibles libremente en GitHub (https://github.com/marcos-diazg/musica).

MuSiCa permite unha caracterización do perfil mutacional somático a nivel de mostra, o que proporciona grandes beneficios no caso de pequenas cohortes e mostras individuais (Blokzijl et al., 2018). Ambos escenarios son comúns no contexto clínico, onde o perfil mutacional de cada paciente debe ser comprobado fronte ao mesmo conxunto de sinaturas mutacionais de referencia (Rosenthal et al., 2016; Baez-Ortega & Gori, 2019). Así, MuSica establécese como unha ferramenta útil para a caracterización de sinaturas mutacionais na práctica clínica, sempre que se dispoña de datos de SNX tanto de ADN xerminal coma tumoral (xa que é necesario para poder identificar as variantes somáticas).

Nos últimos anos desenvolvéronse outras aplicacións web para realizar análises de sinaturas mutacionais (Baez-Ortega & Gori, 2019; Grolleman, Díaz-Gay, et al., 2019; Hanane et al., 2019). Pmsignature foi a primeira ferramenta web que tivo unha interface gráfica, aínda que só permitía o descubrimento de sinaturas mutacionais *de novo* (a través do seu novidoso modelo probabilístico) e non o axuste segundo un conxunto de sinaturas de referencia (Shiraishi et al., 2015). Pola súa banda, a aplicación web MutaGene proporciona un marco computacional para unha caracterización completa das mutacións tumorais e dos procesos mutacionais asociados, o que permite analizar xenes específicos e buscar potenciais mutacións *driver*. Aínda que está centrada na avaliación dos datos de mostras de cancro dispoñibles publicamente, tamén permite realizar análises de axuste de sinaturas mutacionais, pero neste caso só permite a análise das mostras de unha en unha, limitando así a comparación en cohortes de máis de un paciente (Goncearenco et al., 2017). Como no caso anterior, mSignatureDB ofrece a posibilidade tanto de analizar datos procedentes de mostras de tumor dispoñibles publicamente, como de realizar unha análise de sinaturas mutacionais nunha serie de mostras proporcionadas directamente polos usuarios. Esta análise pode ser *de novo*

(utilizando o paquete mutSignatures (Fantini et al., 2018)) ou mediante axuste de sinaturas (a través de deconstructSigs (Rosenthal et al., 2016)), o cal presenta un tempo de computación moi superior ao de MutationalPatterns e, polo tanto, ao de MuSiCa (P.-J. Huang et al., 2018). Por último, Mutalisk é a aplicación web máis completa con respecto á análise mutacional somática a nivel de mostra ata o momento. Ademais da descomposición de sinaturas, Mutalisk proporciona información sobre hipermutación localizada (denominada *kataegis*), nesgo da cadea transcripcional, contido de GCs, tempo de replicación do ADN, modificacións das histonas e hipersensibilidade á DNase I (J. Lee et al., 2018). Respecto ás súas competidoras, MuSiCa presenta funcionalidades exclusivas para a clasificación de mostras, que se pode realizar a través de *clustering* e análise de compoñentes principais, e que podería ter un potencial importante no ámbito clínico. Así, por exemplo, nunha cohorte dun certo subtipo de cancro moi específico e cun fenotipo ben definido, a comparación dos seus perfís de sinaturas mutacionais con outros de pacientes doutras tipoloxías de cancro podería proporcionar novos coñecementos sobre o defecto xenético responsable. Esta estratexia foi utilizada recentemente con éxito no caso da deficiencia de *NTHL1* e a súa asociación coa sinatura SBS30 (Grolleman, de Voer, et al., 2019).

Como medida da potencial aplicabilidade de MuSiCa, realizouse a replicación da caracterización dos perfís mutacionais somáticos dos tumores de colon procedentes do proxecto TCGA (Muzny et al., 2012). Utilizáronse un total de 433 mostras e conseguiuse reproducir satisfactoriamente as vías moleculares de IMS (dominada polas sinaturas asociadas a un MMR defectuoso: SBS6, SBS15, SBS20 e SBS26), deficiencia no sistema de reparación por corrección das polimerases (ligada á sinatura asociada a mutacións en *POLE*: SBS10) e INC (que se atopou dominada pola sinatura asociada á idade SBS1). Isto último débese a que as alteracións *driver* nesta vía son principalmente de número de copia, mentres que a análise de sinaturas unicamente considera as SNVs, que neste caso serían eventos *passenger* relacionados co proceso de envellecemento.

Por outra banda, no segundo estudo desta tese de doutoramento desenvolveuse e aplicouse unha análise integrada de datos de SEC xerminal e tumoral nunha cohorte de 18 pacientes non relacionados de CCR familiar, xunto cunha caracterización dos perfís mutacionais somáticos realizada coa aplicación MuSiCa desenvolvida anteriormente, co obxectivo de atopar novos xenes candidatos responsables da predisposición xerminal a esta neoplasia.

As mostras utilizadas neste estudo pertencen a unha cohorte máis ampla de CCR familiar (71 pacientes de 38 familias), da que se dispón de datos de SEC xerminal e que foi utilizada previamente en diversos estudos do grupo de investigación (Esteban-Jurado et al., 2015, 2016; Franch-Expósito et al., 2018). Estas familias foron seleccionadas por teren unha agregación forte para a enfermidade, así como por non

presentar defectos xerminais nos xenes de predisposición xa coñecidos. A posibilidade de dispoñer de datos de secuenciación combinados xerminais e tumorais proporcionou, por primeira vez no noso grupo de investigación, a oportunidade de analizar o perfil de alteracións xenéticas somáticas. Neste sentido, a experiencia acumulada na identificación e análise de diferentes tipos de variantes potencialmente patoxénicas nos datos de SEC xerminal (incluíndo SNVs, indels e CNVs) foi explotada e trasladada ao ámbito somático.

Despois da identificación de variantes a través de diferentes softwares (GATK HaplotypeCaller para SNVs/indels xerminais, CoNIFER e ExomeDepth para CNVs xerminais, MuTect2 para SNVs/indels somáticas e ALFRED para predicir LOHs somáticas), empregouse unha análise integrada xerminal-tumoral baseada na hipótese dos dous *hits* de Knudson para a priorización dos xenes máis interesantes como candidatos á predisposición ao CCR. Así, estes XSTs candidatos debían presentar unha alteración xerminal e outra somática de forma que se perdese completamente a súa función. Esta estratexia tamén foi utilizada nalgúns estudos recentes. No caso de Spier e colaboradores, empregaron esta estratexia nunha cohorte de 7 pacientes con polipose adenomatosa, aínda que non puideron identificar ningún xene candidato que seguise o modelo dos dous *hits* (Spier et al., 2016). Por outra banda, nunha análise de máis de 10.000 mostras de diferentes tipos de cancro accesibles publicamente, detectáronse un total de 13 xenes, incluíndo xenes de predisposición a diferentes neoplasias xa coñecidos, como *BRCA1*, *BRCA2* e *ATM*, pero tamén novos potenciais candidatos como é o caso da histona metiltransferase *NSD1* (Park et al., 2018).

Aínda que a análise realizada no noso estudo ten en conta diferentes tipos de variantes, outras posibles alteracións poderían actuar tamén como o primeiro ou segundo *hit* no modelo de Knudson, incluíndo alteracións epixenéticas, como modificacións das histonas ou ARNs non codificantes (microARNs ou ARNs non codificantes longos) (Okugawa et al., 2015), así como defectos en rexións non codificantes do xenoma (que non puideron ser avaliadas por seren os datos de SNX de partida procedentes de SEC). Por outra banda, a estratexia de priorización escollida limita, á súa vez, a selección de candidatos, xa que mecanismos como a haploinsuficiencia fan prescindible o segundo *hit* somático para que o xene afectado a nivel xerminal teña influencia na predisposición hereditaria (Deutschbauer et al. , 2005), como se viu no caso dos xenes *BUB1* e *BUB3* (De Voer et al., 2013). Ademais, poderían usarse outras estratexias de priorización, como a replicación en cohortes adicionais ou os estudos funcionais, que tiveron éxito no caso de xenes candidatos como *RPS20* (Nieminen et al., 2014), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *SETD6* (Martín-Morales et al., 2017) ou *BRF1* (Bellido et al., 2018).

Despois de levar a cabo un control de calidade, que descartou dúas das 18 mostras debido a unha baixa calidade de secuenciación tumoral, aplicáronse as correspondentes *pipelines* de análise xerminal e somático. Na cohorte final de 16 mostras, atopáronse 494 SNVs e 42 indels xerminais, mentres que non se identificou ningunha CNV xerminal que afectase a xenes cunha función compatible coa predisposición ao CCR familiar. Tendo en conta as alteracións somáticas, detectáronse un total de 143 xenes con variantes tanto no xenoma xerminal como no somático (**Figura 21**). En tres destes xenes, *ADCY8*, *HSPG2* e *TTN*, identificáronse dúas SNVs, unha xerminal e outra somática, aínda que o xene *TTN* foi descartado debido á súa gran lonxitude (o que podería provocar a acumulación de variantes simplemente por azar) (Chauveau et al., 2014). Por outra banda, en 141 xenes identificouse unha SNV ou indel xerminal e predíxose unha LOH tumoral como segundo *hit* (tamén incluíndo ao mencionado *HSPG2*). Para reducir o número de xenes a unha primeira selección de 16 candidatos potenciais, foi necesario un proceso de curado manual segundo a información funcional publicada previamente para cada xene (**Figura 21**). Cómpre salientar que se atopou un enriquecemento da reparación do ADN entre as funcións asociadas aos xenes seleccionados (7 de 16 xenes implicados, incluíndo *BLM*, *BRCA2*, *ERCC2*, *PARP2*, *RECQL*, *REV3L* e *RIF1*), o que está de acordo con parte dos xenes hereditarios clásicos de CCR (Valle, Vilar, et al., 2019). Tamén se destacaron aqueles xenes que causan unha síndrome de predisposición ao cancro cando están mutados xerminalmente (*BLM*, *BRCA2*, *ERCC2* e *SMARCA4*) (Rahman, 2014), así como dous xenes asociados a síndromes de predisposición ao CCR coñecidos, o síndrome de Cowden e o de Peutz-Jeghers, que foron detectados nunha mostra cun fenotipo ultrahipermutado a nivel somático (*SEC23B* e *STK11IP*) (D. P. Smith et al., 2001; Yehia et al., 2015). Así, un total de 10 xenes con SNV/indel xerminal e LOH somática foron priorizados por estas dúas estratexias, reparación do ADN e síndromes de cancro hereditario, que xunto cos dous xenes con SNV xerminal e SNV somática fan un total de 12 candidatos a ter en conta para a predisposición ao CCR familiar (**Figura 21**). Adicionalmente, realizouse unha análise de casos e controis para estes xenes empregando datos de 1.006 casos de CCR de aparición precoz da base de datos CanVar (Chubb, Broderick, Dobbins & Houlston, 2016), así como a base de datos ExAC como controis (Lek et al., 2016), obtendo un enriquecemento en casos para as variantes afectando aos xenes *ADCY8*, *BLM*, *BRCA2*, *ERCC2*, *REV3L*, *RIF1*, *SEC23*, *SMARCA4* e *STK11IP*. Tamén se realizou unha caracterización mutacional somática a través da aplicación MuSiCa desenvolvida no primeiro estudo desta tese (Díaz-Gay et al., 2018), co obxectivo de engadir máis evidencia de cara á priorización de candidatos á predisposición ao CCR familiar. Avaliáronse tanto a TMB como as achegas das sinaturas mutacionais segundo as sinaturas de referencia v2 de COSMIC (Wellcome Trust Sanger Institute, 2019a). Cómpre sinalar que se atoparon un total de cinco tumores hipermutados, o que está de

acordo co enriquecemento atopado anteriormente en funcións relacionadas coa reparación do ADN entre os candidatos seleccionados (Campbell et al., 2017).

Con respecto aos 12 candidatos seleccionados inicialmente, os dous que presentaban SNV xerminal e somática (*ADCY8* e *HSPG2*) foron descartados para análises posteriores ao identificarse un rol potencialmente oncoxénico despois dun curado funcional máis completo (mentres que o modelo de Knudson está baseado en XSTs) (B. Sharma et al., 1998; Hong et al., 2013). Por outra banda, entre os xenes con SNV/indel xerminal e inactivación somática predita por LOH, finalmente destacáronse seis xenes, incluídos os xenes de predisposición xa coñecidos para outras neoplasias *BLM*, *BRCA2* e *ERCC2*, así como os xenes asociados á reparación do ADN *RECQL*, *REV3L* e *RIF1*.

BLM e *RECQL* pertencen á familia RecQ de helicases, responsable da apertura do ADN de dobre cadea e con funcións na replicación, recombinación, transcrición e reparación do ADN (Croteau et al., 2014). Cómpre destacar que mutacións xerminais bialélicas en *BLM* causan a síndrome de cancro hereditario de Bloom (Ellis et al., 1995), mentres que no caso de *RECQL*, as variantes atopadas pertencen a un paciente no que se atopou un fenotipo hipermutado (preto de 100 mutacións por megabase secuenciada) no tumor. Ambos os dous xenes tamén foron propostos recentemente como xenes de predisposición ao cancro de mama (Thompson et al., 2012; Cybulski et al., 2015), mentres que *BLM* xa fora proposto previamente para a predisposición ao CCR (de Voer et al. , 2015). *BRCA2* constitúe un dos xenes hereditarios clásicos para o cancro de mama e ovario (Wooster et al., 1995), e no caso da nosa cohorte atopouse a súa dobre alteración xerminal-somática nun paciente pertencente a unha familia con varios membros tamén afectados por cancro de mama. Así, foi seleccionado coma o xene responsable do fenotipo na familia, descartando así a *PARP2*, que fora detectado no mesmo paciente. Respecto a *ERCC2*, as alteracións xerminais bialélicas causan *xeroderma pigmentosum*, unha síndrome hereditaria responsable dunha susceptibilidade incrementada ao cancro de pel (Frederick et al., 1994). Este xene, pertencente á vía de reparación do ADN por excisión de nucleótidos, tamén foi proposto como candidato á predisposición ao cancro de mama e ovario (Rump et al., 2016). Finalmente, *REV3L* e *RIF1* asociáronse coa reparación do ADN en dúas vías diferenciadas, a síntese de ADN translesión e a reparación de roturas de ADN de dobre cadea por unión de extremos non homólogos, respectivamente (Lange et al., 2011; Escribano-Díaz et al., 2013). Ademais, tamén se descartou o xene candidato *SMARCA4* (cuxa inactivación fora predita na mesma familia que *REV3L*), despois de avaliar a validación de LOH mediante secuenciación Sanger realizada en estudos anteriores (Esteban-Jurado et al., 2015, 2016). Tamén cómpre destacar que os xenes *SEC23B* e *STK11IP*, detectados nunha mostra cun tumor ultrahipermutado (máis de 500 mutacións por megabase), finalmente foron descartados posto que se esperaba un

defecto nalgunha das vías de reparación do ADN neste caso debido ao elevado número de mutacións atopado.

En canto á análise de sinaturas mutacionais, atopouse un predominio da sinatura SBS1 asociada á idade, o que está de acordo coas análises anteriores realizadas con MuSiCa na cohorte de cancro de colon do TCGA, para mostras sen IMS nin mutacións en *POLE*. Non obstante, isto estaría en desacordo cos altos valores de TMB atopados na cohorte, especialmente nos cinco casos hipermutados (Muzny et al., 2012). Tamén hai que sinalar que non se atopou ningunha das sinaturas asociadas a defectos na reparación do ADN cunha contribución significativa no perfil mutacional das mostras analizadas.

A análise integrada xerminal-tumoral desenvolvida está de acordo coas recentes recomendacións do Clinical Genome Resource, que propón o uso da TMB e a análise de sinaturas na práctica clínica rutineira. Tamén considera a avaliación do segundo *hit* somático, aínda que neste caso se recomenda unha análise caso por caso e baixo o asesoramento dun panel multidisciplinario de expertos en cada centro (Walsh et al., 2018). Cómpre sinalar que os potenciais candidatos á predisposición xerminal ao CCR familiar identificados neste estudo poderían ser útiles nun futuro na práctica clínica, permitindo mellorar o diagnóstico nas familias afectadas. Non obstante, a validación das alteracións xenéticas atopadas mediante técnicas ortogonais, así como a replicación en cohortes independentes de CCR familiar e estudos funcionais serían necesarios para a confirmación da súa asociación co CCR hereditario, así como para proporcionar novos coñecementos sobre os mecanismos moleculares implicados.

Conclusións

1. *Mutational Signatures in Cancer* (MuSiCa) é unha aplicación web sinxela e de acceso libre desenvolvida a través da plataforma Shiny para realizar a caracterización mutacional somática de mostras de cancro.

2. MuSiCa estableceuse como unha das aplicacións web de referencia para o cálculo da carga mutacional tumoral e a caracterización de sinaturas mutacionais segundo as sinaturas de referencia de COSMIC, sendo amplamente utilizada desde a súa publicación.

3. A clasificación de mostras por *clustering* e análise de compoñentes principais segundo as achegas das diferentes sinaturas mutacionais é unha característica distintiva de MuSiCa, que non está dispoñible en ningunha das aplicacións competidoras que existen para realizar análise de sinaturas mutacionais.

4. A caracterización molecular de mostras somáticas de CCR procedentes do proxecto TCGA replicouse dunha forma sinxela e precisa mediante a análise de sinaturas mutacionais de MuSiCa.

5. A análise integrada de datos de SEC xerminais e tumorais, tendo en conta diferentes clases de variantes xenéticas e baseada na hipótese clásica dos dous *hits* de Knudson e a caracterización mutacional somática, demostrouse útil para a identificación de novos XSTs candidatos a estar involucrados na predisposición ao CCR familiar.

6. Identificáronse seis xenes como potenciais candidatos á predisposición xerminal ao CCR familiar, incluíndo xenes coñecidos pola súa implicación na predisposición a outros neoplasias, como é o caso de *BLM*, *BRCA2* e *ERCC2*, así como xenes asociados á reparación do ADN, *RECQL*, *REV3L* e *RIF1*.

7. A análise do perfil mutacional somático pode ser útil no descubrimento do defecto xerminal responsable. No noso estudo, isto foi exemplificado por un xene candidato ligado á reparación do ADN, *RECQL*, que se atopou mutado no ADN xerminal dun paciente cun fenotipo hipermutado no tumor, reforzando o papel potencial deste xene no CCR hereditario.

Summary in Catalan | Resum en català

Identificació de nous gens candidats per a la predisposició germinal al càncer colorectal familiar a través de la caracterització mutacional somàtica

Introducció

El càncer colorectal (CCR) és una de les neoplàsies malignes més comuns i amb més mortalitat associada al món, amb més d'un milió i mig de nous casos i més de 800.000 morts cada any (**Figura 1**) (Bray et al., 2018). La major incidència es troba en les regions més desenvolupades, entre elles Austràlia, Nova Zelanda, Europa, Àsia oriental i Amèrica del Nord (**Figura 2**). A Europa, el CCR representa el segon tipus de càncer en incidència i mortalitat considerant ambdós sexes, mentre que a Espanya es tracta del tipus de càncer amb major incidència i el segon darrere del càncer de pulmó en mortalitat (Ferlay et al., 2019). Com a malaltia complexa, l'etiologia del CCR implica la combinació de diferents factors de risc. A més de factors no modificables, com l'edat o el sexe masculí, els factors ambientals han estat associats a un augment en la incidència de CCR, particularment amb l'anomenada occidentalització de la dieta i l'estil de vida (Brenner et al., 2014).

El CCR va ser un dels primers tumors sòlids caracteritzats a nivell molecular, amb diferents vies de senyalització implicades en l'inici i la progressió de la carcinogènesi (Fearon, 2011). Aquest procés es va descriure inicialment a través de la seqüència adenoma-carcinoma, on una acumulació d'alteracions genètiques en oncogens i gens supressors de tumors (GSTs) dóna lloc a una transició des d'una lesió precursora (anomenada pòlip o adenoma) a un carcinoma, a través de diferents estats intermedis caracteritzats per alteracions genètiques i/o epigenètiques específiques (**Figura 3**) (Vogelstein et al., 1988; Kuipers et al., 2015). Els oncogens es defineixen com aquells gens on l'activació accelera el desenvolupament tumoral, mentre que en el cas dels GSTs és la seva pèrdua d'expressió la que està lligada a l'adquisició del fenotip neoplàsic (Bashyam et al., 2019). Aquest fenotip es caracteritza principalment per un creixement cel·lular descontrolat i la supressió dels mecanismes de mort i reparació cel·lulars, així com per l'adquisició de les capacitats d'invasió i metàstasi (**Figura 4**) (Hanahan & Weinberg, 2000, 2011). El defecte molecular inicial en la majoria dels tumors colorectals (més d'un 70%) succeeix en el GST *APC*, defecte que provoca la desregulació de la via de senyalització Wnt/β-catenina (Kinzler & Vogelstein, 1996; Brenner et al., 2014), tot i que altres vies de senyalització es poden veure també afectades durant la transformació neoplàsica, incloent RAS-RAF-MAPK, PI3K-AKT, TGFβ i p53 (Kuipers et al., 2015). Recentment s'ha identificat una via de carcinogènesi colorectal alternativa, iniciada per

una tipologia de lesions precanceroses diferenciada, les lesions serrades, que actualment es coneix que representen més del 15% dels casos de CCR i que tenem característiques histològiques i moleculars diferenciades respecte els adenomes convencionals (**Figura 3**) (Carballal et al., 2013; IJspeert et al., 2015).

A nivell molecular es consideren tres vies principals per a la carcinogènesi colorectal: inestabilitat cromosòmica (INC), inestabilitat de microsatèl·lits (IMS) i la caracteritzada per un fenotip de hipermetilació d'illes de dinucleòtids CpG (CIMP, per les seves sigles en anglès *CpG island methylator phenotype*) (**Figura 5**). La INC, caracteritzada per l'acumulació d'alteracions del nombre de còpia, va ser la primera via molecular descrita i és l'origen de la majoria de casos de CCR, especialment dels esporàdics (fins a un 85% d'aquests últims). La IMS es defineix per alteracions en els microsatèl·lits (seqüències repetitives d'ADN localitzades al llarg del genoma), que apareixen en forma de petites insercions o delecions (indels), donant lloc a mutacions de terminació de la proteïna per canvi en la pauta de lectura. Aquestes mutacions haurien de ser corregides pel sistema de reparació de l'ADN denominat reparació de mal aparellament de bases (MMR, de l'anglès *mismatch repair*). Quan aquest sistema no funciona correctament, apareix el fenotip d'IMS, àmpliament utilitzat com a biomarcador per a la detecció d'un MMR deficient en CCR i lligat a hipermutació. Per la seva banda, el CIMP es lliga a la hipermetilació dels promotors de nombrosos GSTs associats al càncer, el que provoca la supressió de la seva transcripció (Carethers & Jung, 2015; Kuipers et al., 2015). Recentment s'ha descrit una nova classificació molecular per al CCR basada en patrons d'expressió gènica, els anomenats subtipus moleculars consens (**Figura 6**) (Guinney et al., 2015; Dienstmann et al., 2017).

La predisposició germinal a malalties complexes, com és el cas del CCR, implica una distribució diversa de variants genètiques, que poden ser classificades segons la seva freqüència en la població, o bé respecte el risc associat a desenvolupar una determinada malaltia (conegut com penetrància) (**Figura 7**) (McCarthy et al., 2008; Manolio et al., 2009). Les variants d'alta penetrància es defineixen com aquelles que causen un major efecte en la susceptibilitat a la malaltia, però que són menys freqüents en la població. S'han lligat a malalties que segueixen un patró mendelià (Mendel, 1866), on l'alteració d'un únic gen és freqüentment la responsable del fenotip. Aquestes variants s'han identificat clàssicament a través d'estudis de lligament, també en el cas de les síndromes hereditàries de predisposició al CCR (**Figura 8**) (Bodmer et al., 1987; Lindblom et al., 1993; Peltomaki et al., 1993). D'altra banda, les variants de baixa penetrància es caracteritzen per ser comuns en la població general i tenir un menor efecte a nivell individual en el desenvolupament de la malaltia. No obstant això, una combinació d'aquestes variants, juntament amb la interacció amb factors de risc ambientals, pot contribuir significativament a la predisposició a la malaltia. S'han detectat majoritàriament per estudis d'associació del genoma complet (GWAS, de

l'anglès *genome wide association studies*), que en cas del CCR han permès identificar al voltant de 130 variants implicades i que expliquen un 7-8% de la susceptibilitat associada a aquesta malaltia (**Figura 8**) (Jiao et al., 2014; Peters et al., 2015; Buniello et al., 2019). En determinades malalties, com és el cas del CCR, la ràtio d'heretabilitat estimada pel que fa als estudis clàssics en bessons i famílies (12-35%) no està d'acord amb l'heretabilitat explicada per les variants genètiques amb una associació coneguda amb la malaltia (2-8%), de manera que això comporta una heretabilitat *no filiada* (Jiao et al., 2014; Valle, Vilar, et al., 2019). Aquesta heretabilitat estaria relacionada, en part, amb aquelles variants no prou freqüents per a ser identificades per GWAS ni amb un efecte en el desenvolupament de la malaltia suficient per a ser detectades per estudis familiars de lligament (**Figures 7-8**) (Manolio et al., 2009). En aquest sentit, la seqüenciació de nova generació (SNG) s'ha desmarcat com l'eina més utilitzada per a la identificació d'aquestes variants. Aquesta tècnica ha revolucionat el camp de la genètica, ja que permet la identificació de diferents classes de variants implicades en la predisposició a diferents malalties a un baix cost relatiu, incloent principalment variants d'un únic nucleòtid (SNVs, de l'anglès *single nucleotide variants*) i indels, però també variants de nombre de còpia (CNVs, de l'anglès *copy number variants*) (Lappalainen et al., 2019). Les CNVs es defineixen com fragments d'ADN d'una mida superior a 50 nucleòtids amb variacions en el nombre de còpia (delecions o duplicacions) respecte al genoma de referència (Alkan et al., 2011). L'aplicació de la SNG més reeixida en els estudis biomèdics traslacionals ha estat la seqüenciació de l'exoma complet (SEC), és a dir, de totes les regions codificants del genoma (Teer & Mullikin, 2010). No obstant això, de cara a la identificació de nous gens de predisposició, aquesta tecnologia necessita la implementació d'una estratègia de priorització, que possibiliti reduir l'alt nombre de variants que s'identifiquen inicialment (**Figura 9**) (Ott et al., 2015).

Les síndromes hereditàries de predisposició al CCR relacionades amb variants genètiques d'alta penetrància representen el 2-8% de tots els casos, i fins al 6-10% si es consideren també les variants de penetrància moderada. Diferents gens, que formen part de diferents vies de senyalització, han estat implicats en aquestes síndromes, caracteritzades per estar originades per diferents tipologies de lesions preneoplàsiques (o pòlips) (**Figura 10**) (Tomlinson, 2015). Aquestes síndromes es classifiquen fenotípicament segons la presència o no d'una acumulació d'aquestes lesions precursores anomenada poliposi (**Figura 11**) (Valle, Vilar, et al., 2019). Les síndromes polipòsiques es divideixen al seu torn segons el tipus de pòlips trobats en els pacients. Amb poliposi adenomatosa estan la poliposi adenomatosa familiar i la seva variant atenuada (lligades fonamentalment a mutacions germinals en el gen *APC*) (Leppert et al., 1987, 1990), la poliposi associada a *MUTYH* (Al-Tassan et al., 2002), la poliposi associada a la reparació de l'ADN per correcció realitzada per les polimerases (lligada a defectes germinals en *POLE* i *POLD1*) (Palles et al., 2013) i la síndrome tumoral associada

a *NTHL1* (implicat en fins a 14 tipus tumorals diferents) (Weren , Ligtenberg, et al., 2015). Per la seva banda, originada per pòlips serrats, apareix la síndrome de poliposi serrada (de la que només s'ha proposat un gen candidat per la seva predisposició hereditària, *RNF43*, encara que amb controvèrsia) (Gala et al., 2014); sorgint de pòlips hamartomatosos, la síndrome de Peutz-Jeghers (lligat a defectes germinals en *STK11*) (Giardiello et al., 1987), la síndrome de poliposi juvenil (*BMPR1A*, *SMAD4*) (Howe et al., 1998, 2001) i la síndrome tumoral *PTEN*-hamartoma / síndrome de Cowden (*PTEN*) (Liaw et al., 1997); i a través d'una combinació de les tres tipologies de pòlips, la síndrome hereditària de poliposi mixta (*GREM1*) (Jaeger et al., 2012). D'altra banda, de les síndromes no polipòsiques destaca la síndrome de Lynch, associada a mutacions germinals en els gens del sistema de MMR (*MLH1*, *MSH2*, *MSH6*, *PMS2*) i que constitueix la síndrome de CCR hereditari més freqüent (H. T. Lynch et al., 2015). A causa d'això últim, s'han desenvolupat una sèrie de guies clíniques per a la identificació de les famílies amb més probabilitat de ser portadores d'aquesta síndrome (**Figura 12**) (Vasena et al., 1999; Umar et al., 2004). Les esmentades síndromes presenten en general una herència autosòmica dominant, excepte en el cas d'aquelles lligades a mutacions en gens de la via de reparació de l'ADN per escissió de bases (BER, de l'anglès *base excision repair*), *MUTYH* i *NTHL1*, el patró d'herència dels quals és autosòmic recessiu (Valle, Vilar, et al., 2019).

A més de les comentades síndromes de predisposició hereditàries (que expliquen fins a un 8% de l'heretabilitat), s'especula que els factors genètics estiguin darrere d'un 12-35% del total de casos de CCR (Lichtenstein et al., 2000; Jiao et al., 2014; Peters et al., 2015). Aquesta heretabilitat *no filiada* ha estat objecte d'estudi en els darrers anys amb l'objectiu d'identificar nous gens candidats, que podrien tenir un fort impacte en el consell genètic en les famílies afectades. La SNG ha estat la tecnologia principalment utilitzada en aquest esforç d'identificació de nous gens implicats en la predisposició al CCR (Valle, de Voer, et al., 2019). Així, un gran nombre de gens candidats ha estat proposat per diferents grups de recerca, incloent *BUB1*, *BUB3* (de Voer et al., 2013), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *BLM* (de Voer et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *MIA3* (Schubert et al., 2017), *SETD6* (Martín-Morales et al., 2017) i *BRF1* (Bellido et al., 2018) com els més prometedors en base als estudis funcionals realitzats i a la validació en cohorts de CCR familiar addicionals.

Segons la hipòtesi dels dos *hits* de Knudson, el desenvolupament neoplàsic comença amb dos esdeveniments mutacionals en un únic gen (un GST), que impedeixen la seva expressió. Així, les diferències observades entre les formes hereditàries i esporàdiques/no hereditàries d'un determinat càncer es deuen a la diferent combinació d'aquestes alteracions genètiques o *hits* (que poden ser de diferents classes: SNVs, indels, CNVs, pèrdues d'heterozigositat (LOH, l'anglès *loss of heterozygosity*) o

alteracions en la metilació). En el cas d'un càncer hereditari existiria una primera alteració en l'ADN germinal seguida per un segon *hit* somàtic, mentre que en els casos esporàdics es trobarien directament dues mutacions en les cèl·lules tumorals (**Figura 13**). Així, s'explicaria l'aparició més primerenca dels càncers hereditaris, ja que només és necessari un esdeveniment mutacional en el tumor per al desenvolupament de la malaltia (Knudson, 1971).

Tots els càncers es caracteritzen per múltiples mutacions somàtiques. Aquestes es classifiquen en mutacions *driver* o *passenger* segons els seus efectes en el desenvolupament tumoral (Stratton et al., 2009). Tot i que la identificació de mutacions *driver* s'ha prioritzat en la majoria dels estudis de seqüenciació, degut al fet que es seleccionen positivament i estan darrere del desenvolupament carcinogènic, les mutacions passenger també han demostrat ser informatives. De fet, el nombre total de mutacions acumulades per un tumor (denominat com a càrrega mutacional tumoral (TMB, de l'anglès *tumor mutational burden*)), altament variable entre tipus tumorals i també dins del mateix càncer (**Figura 14**) (Alexandrov, Nik-Zainal , Wedge, Aparicio, et al., 2013), s'ha postulat en els últims anys com un prometedor biomarcador per inmunoteràpies, per la seva relació amb la càrrega de neoantígens (Chalmers et al., 2017).

A més de la caracterització de la TMB, les mutacions passenger també són responsables de l'aparició d'un nou camp d'estudi en els últims anys. Assumint que els patrons d'aquestes mutacions no varien amb el temps, poden ser utilitzades com una imatge representativa dels mecanismes mutacionals que han estat actius durant el procés carcinogènic (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Cada procés mutacional deixa una empremta particular en el genoma d'una cèl·lula, un perfil de mutacions específic denominat signatura mutacional. Mecanismes cel·lulars endògens, com la replicació i la reparació de l'ADN, poden generar mutacions a causa de la seva taxa d'error intrínseca. D'altra banda, les mutacions també poden ser degudes a exposicions mutagèniques exògenes, com seria el cas del tabac o la llum ultraviolada. Així, el conjunt final de mutacions recollit en un tumor està determinat per la intensitat i la durada de tots els processos mutacionals actius durant el desenvolupament neoplàsic (**Figura 15**) (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

El dany a l'ADN pot aparèixer en forma de diferents tipus de variants genètiques, encara que per la descripció de les signatures mutacionals s'han utilitzat fins al moment principalment les SNVs per raons tècniques. Així, en el conjunt actual de signatures mutacionals de referència es consideren sis tipus de canvi de nucleòtid, segons la pirimidina mutada de la parella de bases de Watson-Crick, incloent quatre possibles transversions, C>A, C>G, T>A i T>G, i dues transicions, C>T i T>C. Per a una

caracterització més estricta dels processos mutacionals responsables de les mutacions, es tenen en compte també les bases adjacents al canvi en els contextos 5 ' i 3', donant lloc a un total de 96 possibilitats (6 substitucions de bases * 4 nucleòtids anteriors * 4 nucleòtids posteriors) (**Figura 16**). D'aquesta manera, cada signatura mutacional està formada per una distribució única d'aquests 96 possibles tipus de mutacions (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). En els últims anys s'ha generat un model matemàtic que ha permès la detecció i quantificació precisa de cadascuna de les signatures mutacionals associades als diferents processos mutagènics implicats en el càncer. Amb aquesta finalitat, es va utilitzar inicialment un algoritme basat en la factorització matricial no negativa denominat SigProfiler, que va ser implementat utilitzant MATLAB (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Recentment, aquest algoritme ha estat traduït a altres llenguatges de programació oberts (Gehring et al., 2015; Blokzijl et al., 2018) i addicionalment han sorgit noves estratègies computacionals per a la identificació de signatures mutacionals (Kasar et al., 2015; Shiraishi et al., 2015; Baez-Ortega & Gori, 2019).

A través d'aquests models computacionals ha estat possible l'extracció de signatures mutacionals de referència, cadascuna d'elles associada a un procés mutagènic específic del qual, en alguns casos, s'ha pogut identificar la seva etiologia. Aquestes signatures de referència permeten que l'anàlisi de signatures mutacionals no només es restringeixi a la identificació agnòstica de noves signatures (conegut com a anàlisi *de novo*, és a dir, sense utilitzar cap coneixement previ), sinó que també fan possible la caracterització dels processos mutacionals implicats a nivell de mostra respecte a una referència (denominada com anàlisi d'ajust de signatures mutacionals). Per dur a terme aquest tipus d'anàlisi, més orientada a la seva aplicació en la pràctica clínica, també han sorgit noves eines bioinformàtiques en els darrers anys (Rosenthal et al., 2016; Blokzijl et al., 2018). Malauradament, encara estan orientades a experts bioinformàtics, de manera que romanen inaccessibles per a una part important de la comunitat científica. El nombre de signatures de referència ha anat creixent de mica en mica, a mesura que el nombre de mostres tumorals analitzades s'ha anat incrementant, degut al successiu augment de potència estadística del model matemàtic. En una primera aplicació d'aquesta metodologia es van extreure cinc signatures mutacionals de SNVs (posteriorment reduïdes a quatre després d'una optimització del model) d'una cohort de 21 mostres de càncer de mama (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). A continuació, el conjunt de referència de signatures mutacionals es va ampliar a 21 (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013), i posteriorment a 30 després de l'aplicació d'aquesta metodologia a unes 12.000 mostres de 40 tipologies diferents de càncer (**Figura 17**) (Alexandrov et al., 2015; Tate et al., 2018). Aquest conjunt de 30 signatures mutacionals de referència s'ha utilitzat en la gran majoria de publicacions que han realitzat anàlisi de signatures

mutacionals fins ara (Grolleman, Díaz-Gay, et al., 2019), i es pot trobar com a part de la base de dades COSMIC (versió 2 – marzo 2015) (Wellcome Trust Sanger Institute, 2019a). Finalment, l'actual conjunt de signatures mutacionals de referència s'ha extret de més de 23.000 mostres de càncer i es compon de 49 signatures de SNVs (també anomenades signatures de SBSs, de l'anglès *single base substitutions*), mentre que ha incorporat també altres tipologies de variants, incloent 17 signatures associades a indels i 11 lligades a substitucions de dues bases consecutives (**Figura 18**) (Alexandrov et al., 2019). També es troba disponible a COSMIC (versió 3 – maig 2019) (Wellcome Trust Sanger Institute, 2019b).

S'ha trobat una distribució diferent de signatures mutacionals en els diferents teixits, el que concorda amb les diferents ràtios de reemplaçament cel·lular, així com amb la diferent influència de les exposicions ambientals segons el teixit en qüestió. Algunes signatures, com és el cas de SBS1, SBS5 i SBS40, s'han associat amb l'edat de diagnòstic, reflectint la influència del procés d'envelliment en la carcinogènesi (Alexandrov et al., 2015, 2019). Pel que fa al CCR, segons la informació disponible a COSMIC i a una sèrie de publicacions recents, hi ha una contribució de diferents signatures mutacionals, incloent les esmentades signatures relacionades amb l'envelliment. Tot i això, aquestes signatures contribueixen a un nombre reduït de mutacions, en comparació amb aquelles relacionades específicament amb dos coneguts defectes moleculars presents en el CCR: les deficiències en els processos de reparació de l'ADN per MMR i per correcció de les polimerases (7 signatures diferenciades en el cas d'un mal funcionament del sistema de MMR: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 i SBS44, i les signatures SBS10a i SBS10b en el cas de mutacions en el domini exonucleasa de la polimerasa *POLE*) (Nagahashi et al., 2016; Alexandrov et al ., 2019). A més, algunes d'aquestes signatures s'han associat a la concurrència d'alteracions genètiques en els dos sistemes de reparació de l'ADN, incloent les signatures SBS14 (mutació en *POLE* i en els gens del MMR) i SBS20 (mutació en *POLD1* i MMR defectuós), sent trobada també aquesta última en casos de CCR (Haradhvala et al., 2018; Alexandrov et al., 2019). Recentment s'ha ampliat l'espectre de signatures mutacionals associades al CCR, amb la inclusió de dues signatures relacionades amb el sistema BER de reparació de l'ADN i específicament amb defectes en dos coneguts gens de predisposició al CCR: *MUTYH* (signatura SBS36) (Pilati et al ., 2017; Viel et al., 2017) i *NTHL1* (SBS30) (Drost et al., 2017; Grolleman, de Voer, et al., 2019). Altres signatures també han estat associades al CCR, encara que amb un paper menys prevalent, com és el cas de SBS2, SBS13 (lligades a l'activitat de les desaminases APOBEC), SBS3 (sistema de reparació de l'ADN per recombinació homòloga defectuós i mutacions en *BRCA1/2*), SBS9 (activitat de la polimerasa eta), SBS17a, SBS17b, SBS18 (dany a l'ADN provocat per les espècies reactives de l'oxigen), SBS12, SBS28, SBS37 i SBS41 (etiologia

desconeguda), així com una nova signatura identificada per Roerink i col·laboradors en un estudi recent (Nagahashi et al., 2016; Roerink et al., 2018; Alexandrov et al., 2019).

Les signatures mutacionals, així com la TMB, poden ser utilitzades per a la identificació dels defectes genètics germinals que han estat actius durant l'origen i l'evolució d'un determinat càncer. Això és particularment evident per a aquelles signatures mutacionals amb una etiologia coneguda i, en particular, per a aquelles associades a processos mutacionals responsables de síndromes hereditàries de predisposició al càncer, com les derivades de defectes en els mecanismes de reparació de l'ADN (**Figura 19**) (J . Ma et al., 2018; Van Hoeck et al., 2019). En el cas dels sistemes de correcció de les polimerases i MMR, els defectes germinals s'han identificat lligats a una TMB alta, és a dir, a tumors hipermutats, i addicionalment a les esmentades signatures mutacionals característiques (Muzny et al., 2012; Kandoth et al., 2013; Alexandrov et al., 2019). Les signatures mutacionals, així com l'anàlisi de la TMB, podrien ajudar en la identificació i el descobriment dels processos mutacionals responsables de les diferents síndromes hereditàries de càncer, així com afavorir el diagnòstic genètic i la selecció de tractaments en els pacients, com s'ha demostrat recentment en el cas de les alteracions en els gens *BRCA1/2* i *NTHL1* (Davies et al., 2017; Grolleman, de Voer, et al., 2019).

Hipòtesi

El CCR és una malaltia complexa i, per tant, amb una etiologia en la qual es barregen factors genètics i ambientals. La predisposició genètica està darrere de fins a un 35% dels CCRs segons estudis familiars i de bessons, mentre que les síndromes de predisposició conegudes i associades a defectes genètics germinals específics només expliquen un 2-8% dels casos. D'aquesta manera, s'observa una heretabilitat *no filiada* per aquesta neoplàsia. La SNG és la tècnica més adequada per dur a terme la identificació de nous gens implicats en la predisposició al CCR, com s'ha demostrat en estudis recents en gens com *POLD1*, *POLE* i *NTHL1*. No obstant això, aquesta tecnologia identifica un gran nombre de variants genètiques en cada pacient, generant així la necessitat d'una estratègia de priorització. En aquest sentit, segons la clàssica hipòtesi dels dos *hits* de Knudson, a més de les alteracions genètiques germinals, també les somàtiques poden jugar un paper fonamental en proporcionar nou coneixement respecte a la predisposició hereditària al CCR. Així doncs, l'anàlisi del perfil mutacional somàtic s'ha utilitzat recentment per a la identificació de nous gens de predisposició al CCR, i com un biomarcador prometedor de cara al diagnòstic, pronòstic i tractament d'aquesta neoplàsia. Tot i que s'han desenvolupat diversos paquets bioinformàtics per realitzar aquest tipus d'anàlisi, encara roman inaccessible per a una proporció substancial de la comunitat científica.

Objectius

L'objectiu principal de la present tesi doctoral és el d'identificar nous gens candidats que puguin estar implicats en la predisposició germinal al CCR familiar. Una anàlisi combinada germinal-tumoral de dades de SEC i una aplicació bioinformàtica per realitzar caracterització mutacional somàtica es desenvoluparan per ser utilitzades com a estratègies de priorització.

Amb aquesta finalitat, es duran a terme els següents objectius específics:

1. Desenvolupament d'una aplicació computacional per realitzar anàlisis dels perfils mutacionals somàtics, a través d'una interfície senzilla adequada per a investigadors no especialitzats en bioinformàtica i accessible lliurement mitjançant una pàgina web. Estaran disponibles tant la caracterització de la TMB com l'ajust de les signatures mutacionals segons les signatures de referència versió 2 de COSMIC, així com la classificació de mostres per clustering i anàlisi de components principals.

2. Anàlisi integrat basat en la hipòtesi dels dos *hits* de Knudson de dades de SEC procedents d'ADN germinal i tumoral d'una cohort de 18 pacients de CCR familiar, amb l'objectiu d'identificar nous GSTs potencials. Es tindran en compte diferents classes d'alteracions genètiques, mentre que els gens candidats es seleccionaran quan tant l'ADN germinal com el tumoral estiguin afectats per una d'aquestes alteracions.

3. Caracterització somàtica mutacional de l'esmentada cohort de CCR familiar utilitzant l'eina bioinformàtica desenvolupada prèviament, a través de l'anàlisi de la càrrega mutacional tumoral i les signatures mutacionals.

Resultats i discussió

El primer dels estudis publicats com a part d'aquesta tesi doctoral presenta el desenvolupament de l'aplicació MuSiCa (de l'anglès *Mutational Signatures in Cancer*), que constitueix una de les primeres eines web disponibles per realitzar una caracterització mutacional somàtica completa dels tumors seqüenciats amb tècniques de SNG.

Tant el càlcul de la TMB com la reconstrucció dels perfils mutacionals somàtics segons les signatures mutacionals de referència versió 2 de COSMIC (Wellcome Trust Sanger Institute, 2019a) estan disponibles a MuSiCa. Una futura actualització de l'aplicació serà necessària de cara a l'adaptació a la nova versió 3 d'aquestes signatures de referència, que haurà d'incloure les noves classes de variants a tenir en compte. Respecte a l'ajust de signatures mutacionals, MuSiCa utilitza com a base el paquet de R/Bioconductor MutationalPatterns (Blokzijl et al., 2018), que es basa en la resolució d'un problema d'optimització de mínims quadrats no negatius a través d'un algoritme de mètode de conjunt actiu (Lawson & Hanson, 1974) inclòs en el paquet de R pracma (Borchers, 2019). MuSiCa proporciona una interfície gràfica a aquest paquet, creada

mitjançant el paquet de R Shiny (W. Chang et al., 2019) i específicament dissenyada per investigadors no especialitzats en bioinformàtica, així com algunes característiques addicionals. MuSiCa està disponible de forma gratuïta com a part de la pàgina web del nostre grup de recerca (http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html), el que permet el seu ús de forma senzilla per part de qualsevol membre de la comunitat científica sense necessitat de grans recursos a nivell de computació. De fet, segons les dades recollides per la plataforma Google Analytics durant els primers 14 mesos des de la publicació de l'article de MuSiCa, 1.344 usuaris únics d'un total de 53 països diferents han accedit al web de l'aplicació, en un total de 3.045 sessions (**Figura 20**). També és possible utilitzar MuSiCa de forma local, per a això les dependències requerides per a la seva instal·lació, així com el codi font en R, estan lliurement disponibles a GitHub (https://github.com/marcos-diazg/musica).

MuSiCa permet una caracterització del perfil mutacional somàtic a nivell de mostra, el que proporciona grans beneficis en el cas de petites cohorts i mostres individuals (Blokzijl et al., 2018). Tots dos escenaris són comuns en l'entorn clínic, on el perfil mutacional de cada pacient hauria de ser contrastat respecte al mateix conjunt de signatures mutacionals de referència (Rosenthal et al., 2016; Baez-Ortega & Gori, 2019). Així, MuSiCa s'estableix com una eina útil per a la caracterització de signatures mutacionals en la pràctica clínica, sempre que es disposi de dades de SNG tant d'ADN germinal com tumoral (ja que és necessari per poder identificar les variants somàtiques).

Altres aplicacions web han estat desenvolupades en els últims anys per realitzar anàlisis de signatures mutacionals (Baez-Ortega & Gori, 2019; Grolleman, Díaz-Gay, et al., 2019; Hanane et al., 2019). Pmsignature va ser la primera eina web que va disposar d'una interfície gràfica, tot i que només permetia el descobriment de signatures mutacionals *de novo* (a través del seu nou model probabilístic) i no l'ajust segons un conjunt de signatures de referència (Shiraishi et al., 2015). Per la seva banda, l'aplicació web MutaGene proporciona un marc computacional per a una completa caracterització de les mutacions tumorals i els processos mutacionals associats, permetent analitzar gens específics i buscar potencials mutacions *driver*. Encara que està enfocada en l'avaluació de dades de mostres de càncer disponibles públicament, també permet realitzar anàlisis d'ajust de signatures mutacionals, però en aquest cas només permet analitzar les mostres d'una en una, limitant així la comparació en cohorts de més d'un pacient (Goncearenco et al., 2017). Igual que en el cas anterior, mSignatureDB proporciona la possibilitat d'analitzar dades de mostres tumorals disponibles públicament, així com de realitzar una anàlisi de signatures mutacionals en una sèrie de mostres proporcionades directament pels usuaris. Aquesta anàlisi pot ser *de novo* (utilitzant el paquet mutSignatures (Fantini et al., 2018)) o mitjançant ajust de signatures (a través de deconstructSigs (Rosenthal et al., 2016)), el qual presenta un temps de computació molt superior al de MutationalPatterns i, per tant, de MuSiCa (P.-

J. Huang et al., 2018). Finalment, Mutalisk constitueix l'aplicació web més completa respecte a l'anàlisi mutacional somàtic a nivell de mostra fins a l'actualitat. A més de la descomposició de signatures, Mutalisk proporciona informació sobre hipermutació localitzada (denominada *kataegis*), biaix de la cadena transcripcional, contingut de GCs, temps de replicació de l'ADN, modificacions de les histones i hipersensibilitat a la DNase I (J. Lee et al ., 2018). Respecte a les seves competidores, MuSiCa presenta funcionalitats exclusives de cara a la classificació de mostres, que es pot realitzar a través de *clustering* i anàlisi de components principals, i que podria tenir un important potencial en l'entorn clínic. Així, per exemple, en una cohort d'un cert subtipus de càncer molt específic i amb un fenotip ben definit, la comparació dels seus perfils de signatures mutacionals amb altres de pacients d'altres tipologies de càncer podria proporcionar nous coneixements respecte al defecte genètic responsable. Aquesta estratègia ha estat utilitzada satisfactòriament en el cas de la deficiència de *NTHL1* i la seva associació amb la signatura SBS30 (Grolleman, de Voer, et al., 2019).

Com a mesura de la potencial aplicabilitat de MuSiCa es va dur a terme la replicació de la caracterització dels perfils mutacionals somàtics dels tumors de còlon procedents del projecte TCGA (Muzny et al., 2012). Es van utilitzar un total de 433 mostres i es va aconseguir reproduir satisfactòriament les vies moleculars d'IMS (dominada per les signatures associades a un MMR defectuós: SBS6, SBS15, SBS20 i SBS26), deficiència en el sistema de reparació per correcció de les polimerases (lligada a la signatura associada a mutacions en *POLE*: SBS10) i INC (que es va trobar dominada per la signatura associada a l'edat SBS1). Això últim es deu al fet que les alteracions *driver* en aquesta via són principalment de nombre de còpia, mentre que l'anàlisi de signatures únicament considera les SNVs, que en aquest cas serien esdeveniments *passenger* lligats al procés d'envelliment.

D'altra banda, en el segon estudi d'aquesta tesi doctoral s'ha desenvolupat i aplicat una anàlisi integrada de dades de SEC germinal i tumoral en una cohort de 18 pacients no relacionats de CCR familiar, juntament amb una caracterització dels perfils mutacionals somàtics realitzada amb l'aplicació MuSiCa desenvolupada prèviament, amb l'objectiu de trobar nous gens candidats responsables de la predisposició germinal a aquesta neoplàsia.

Les mostres utilitzades en aquest estudi pertanyen a una cohort més àmplia de CCR familiar (71 pacients de 38 famílies), de la qual es disposa de dades de SEC germinal i que ha estat utilitzada prèviament en diversos estudis del grup de recerca (Esteban-Jurado et al., 2015, 2016; Franch-Expósito et al., 2018). Aquestes famílies van ser seleccionades per tenir una forta agregació per a la malaltia, així com per no presentar defectes germinals en els gens de predisposició ja coneguts. La possibilitat de disposar de dades de seqüenciació combinades germinals i tumorals va proporcionar, per primera

vegada al nostre grup de recerca, l'oportunitat d'analitzar el perfil d'alteracions genètiques somàtiques. En aquest sentit, s'ha explotat i traslladat a l'àmbit somàtic l'experiència acumulada en la identificació i l'anàlisi de diferents tipologies de variants potencialment patogèniques en dades de SEC germinal (incloent SNVs, indels i CNVs).

Després de la identificació de variants a través de diferents softwares (GATK HaplotypeCaller per SNVs/indels germinals, CoNIFER i ExomeDepth per CNVs germinals, MuTect2 per SNVs/indels somàtiques i ALFRED per predir LOHs somàtiques), es va utilitzar una anàlisi integrada germinal-tumoral basada en la hipòtesi dels dos *hits* de Knudson per a la priorització dels gens més interessants com a candidats a la predisposició al CCR. Així, aquests GSTs candidats havien de presentar una alteració germinal i una altra somàtica de manera que es perdés completament la seva funció. Aquesta estratègia ha estat també usada en alguns estudis recents. En el cas de Spier i col·laboradors, van utilitzar aquesta estratègia en una cohort de 7 pacients de poliposi adenomatosa, tot i que no van poder identificar cap gen candidat que seguís el model dels dos *hits* (Spier et al., 2016). D'altra banda, en una anàlisi de més de 10.000 mostres de diferents tipus de càncer públicament accessibles, es van detectar un total de 13 gens, incloent gens de predisposició a diferents neoplàsies ja coneguts, com *BRCA1*, *BRCA2* i *ATM*, però també nous potencials candidats com és el cas de la histona metiltransferase *NSD1* (Park et al., 2018).

Tot i que l'anàlisi realitzada en el nostre estudi té en compte diferents tipus de variants, altres possibles alteracions podrien actuar també com a primer o segon *hit* en el model de Knudson, incloent alteracions epigenètiques, com modificacions de les histones o ARNs no codificants (microARNs o ARNs no codificants llargs) (Okugawa et al., 2015), així com defectes en regions no codificants del genoma (que no han pogut ser avaluades per ser les dades de SNG de partida procedents de SEC). D'altra banda, l'estratègia de priorització escollida limita al seu torn la selecció de candidats, ja que mecanismes com la haploinsuficiència fan prescindible el segon *hit* somàtic perquè el gen afectat a nivell germinal tingui una influència en la predisposició hereditària (Deutschbauer et al. , 2005), com s'ha vist en el cas dels gens *BUB1* i *BUB3* (de Voer et al., 2013). A més, es podrien haver utilitzat altres estratègies de priorització, com la replicació en cohorts addicionals o els estudis funcionals, que han resultat satisfactòries en els casos de gens candidats com *RPS20* (Nieminen et al., 2014), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *SETD6* (Martín-Morales et al., 2017) o *BRF1* (Bellido et al., 2018).

Després de la realització d'un control de qualitat, que va descartar dues de les 18 mostres degut a una baixa qualitat de seqüenciació tumoral, es van aplicar les corresponents *pipeline*s d'anàlisi germinal i somàtic. A la cohort final de 16 mostres, es van trobar 494 SNVs i 42 indels germinals, mentre que cap CNV germinal va ser

identificada afectant gens amb una funció compatible amb la predisposició al CCR familiar. Considerant les alteracions somàtiques, es van detectar un total de 143 gens amb variants tant en el genoma germinal com en el somàtic (**Figura 21**). En tres d'aquests gens, *ADCY8*, *HSPG2* i *TTN*, es van identificar dos SNVs, una germinal i una somàtica, encara que el gen *TTN* va ser descartat per la seva gran longitud (que podria causar l'acumulació de variants simplement per atzar) (Chauveau et al., 2014). D'altra banda, en 141 gens es va identificar una SNV o indel germinal i es va predir un LOH tumoral com a segon *hit* (també incloent a l'esmentat *HSPG2*). Per reduir el nombre de gens a una primera selecció de 16 candidats potencials, va ser necessari un procés de curat manual segons la informació funcional prèviament publicada per a cada gen (**Figura 21**). Cal destacar que es va trobar un enriquiment de la reparació de l'ADN entre les funcions associades als gens seleccionats (7 de 16 gens implicats, incloent *BLM*, *BRCA2*, *ERCC2*, *PARP2*, *RECQL*, *REV3L* i *RIF1*), el que està d'acord amb part dels gens hereditaris clàssics de CCR (Valle, Vilar, et al., 2019). També es van destacar aquells gens causants d'una síndrome de predisposició a càncer quan es troben mutats germinalment (*BLM*, *BRCA2*, *ERCC2* i *SMARCA4*) (Rahman, 2014), així com dos gens associats amb síndromes de predisposició al CCR coneguts, la síndrome de Cowden i la de Peutz-Jeghers, que van ser detectats en una mostra amb un fenotip ultrahipermutat a nivell somàtic (*SEC23B* i *STK11IP*) (D. P. Smith et al., 2001; Yehia et al., 2015). Un total de 10 gens amb SNV/indel germinal i LOH somàtica van ser així prioritzats per aquestes dues estratègies, reparació de l'ADN i síndromes de càncer hereditari, que units als dos gens amb SNV germinal i SNV somàtica fan un total de 12 candidats a tenir en compte per la predisposició al CCR familiar (**Figura 21**). Addicionalment es va realitzar una anàlisi de casos i controls per aquests gens utilitzant les dades de 1.006 casos de CCR d'aparició primerenca de la base de dades CanVar (Chubb, Broderick, Dobbins, & Houlston, 2016), així com la base de dades ExAC com a controls (Lek et al., 2016), obtenint un enriquiment en casos per a les variants afectant als gens *ADCY8*, *BLM*, *BRCA2*, *ERCC2*, *REV3L*, *RIF1*, *SEC23*, *SMARCA4* i *STK11IP*. També es va realitzar una caracterització mutacional somàtica a través de l'aplicació MuSiCa desenvolupada en el primer estudi d'aquesta tesi (Díaz-Gay et al., 2018), amb l'objectiu d'afegir més evidència de cara a la priorització de candidats a la predisposició al CCR familiar. Es van avaluar tant la TMB com les contribucions de les signatures mutacionals segons les signatures de referència v2 de COSMIC (Wellcome Trust Sanger Institute, 2019a). Cal destacar que es van trobar un total de cinc tumors hipermutats, el que està d'acord amb l'enriquiment trobat prèviament en funcions relacionades amb la reparació de l'ADN entre els candidats seleccionats (Campbell et al., 2017).

Pel que fa als 12 candidats inicialment seleccionats, els dos que presentaven SNV germinal i somàtica (ADCY8 i HSPG2) van ser descartats per anàlisis posteriors en identificar-se un paper potencialment oncogènic després d'un curat funcional més

exhaustiu (mentre que el model de Knudson està basat en GSTs) (B. Sharma et al., 1998; Hong et al., 2013). D'altra banda, entre els gens amb SNV/indel germinal i inactivació somàtica predita per LOH, finalment es van destacar sis gens, incloent els gens de predisposició ja coneguts per altres neoplàsies *BLM*, *BRCA2* i *ERCC2*, així com els gens associats a la reparació d'ADN *RECQL*, *REV3L* i *RIF1*.

BLM i *RECQL* pertanyen tots dos a la família RecQ de helicasas, responsables de l'obertura de l'ADN de doble cadena i amb funcions en replicació, recombinació, transcripció i reparació de l'ADN (Croteau et al., 2014). Cal destacar que mutacions germinals bialèlicas en *BLM* causen la síndrome de càncer hereditari de Bloom (Ellis et al., 1995), mentre que en el cas de *RECQL*, les variants trobades pertanyen a un pacient en el qual es va trobar un fenotip hipermutat (al voltant de 100 mutacions per megabase seqüenciada) en el tumor. Tots dos gens han estat, a més a més, proposats recentment com a gens de predisposició a càncer de mama (Thompson et al., 2012; Cybulski et al., 2015), mentre que *BLM* ja havia estat proposat prèviament per a la predisposició al CCR (de Voer et al. , 2015). *BRCA2* constitueix un dels gens hereditaris clàssics per càncer de mama i ovari (Wooster et al., 1995), i en el cas de la nostra cohort s'ha trobat la seva doble alteració germinal-somàtica en un pacient que pertany a una família amb diversos membres també afectats per càncer de mama. Així, ha estat seleccionat com el gen responsable del fenotip en la família, descartant per tant a *PARP2*, que havia estat detectat en el mateix pacient. Pel que fa a *ERCC2*, alteracions germinals bialèlicas causen xeroderma pigmentosum, una síndrome hereditària responsable d'una susceptibilitat incrementada al càncer de pell (Frederick et al., 1994). Aquest gen, pertanyent a la via de reparació de l'ADN per escissió de nucleòtids, també s'ha proposat com a candidat per a la predisposició a càncer de mama i ovari (Rump et al., 2016). Finalment, *REV3L* i *RIF1* s'han associat a la reparació de l'ADN en dues vies diferenciades, la de síntesi d'ADN translesió i la reparació de trencaments de doble cadena d'ADN per unió d'extrems no homòlegs, respectivament (Lange et al., 2011; Escribano-Díaz et al., 2013). Addicionalment, es va descartar el gen candidat *SMARCA4* (del qual també s'havia predit la seva inactivació en la mateixa família que *REV3L*), després d'avaluar la validació de LOH per seqüenciació Sanger realitzada en estudis previs (Esteban-Jurado et al., 2015, 2016). També cal destacar que els gens *SEC23B* i *STK11IP*, detectats en una mostra amb un tumor ultrahipermutat (més de 500 mutacions per megabase), van ser finalment descartats ja que s'esperava un defecte en alguna via de reparació de l'ADN en aquest cas a causa de l'alt nombre de mutacions trobat.

Respecte a l'anàlisi de signatures mutacionals, es va trobar una predominança de la signatura SBS1 associada a l'edat, el que està d'acord amb les anàlisis prèvies realitzades amb MuSiCa en la cohort de càncer de còlon del TCGA, per a les mostres sense IMS ni mutacions en *POLE*. Tanmateix, això estaria en desacord amb els alts valors de TMB trobats a la cohort, especialment en els cinc casos hipermutats (Muzny et al.,

2012). Cal destacar també que no es va trobar cap de les signatures associades a defectes en la reparació de l'ADN amb una contribució significativa en el perfil mutacional de les mostres analitzades.

L'anàlisi integrada germinal-tumoral desenvolupada està d'acord amb les recomanacions recents del Clinical Genome Resource, que proposa l'ús de la TMB i l'anàlisi de signatures en la pràctica clínica rutinària. També considera l'avaluació del segon *hit* somàtic, encara que en aquest cas es recomana una anàlisi cas per cas i sota assessorament d'un panell multidisciplinari d'experts en cada centre (Walsh et al., 2018). Cal destacar que els potencials candidats a la predisposició germinal al CCR familiar identificats en aquest estudi podrien ser útils en un futur en la pràctica clínica, permetent millorar el diagnòstic en les famílies afectades. No obstant això, la validació de les alteracions genètiques trobades mitjançant tècniques ortogonals, així com la replicació en cohorts independents de CCR familiar i estudis funcionals serien necessaris per a la confirmació de la seva associació amb el CCR hereditari, així com per proporcionar nous coneixements sobre els mecanismes moleculars implicats.

Conclusions

1. *Mutational Signatures in Cancer* (MuSiCa) és una aplicació web de maneig senzill i accés lliure desenvolupada a través de la plataforma Shiny per realitzar caracterització mutacional somàtica de mostres de càncer.

2. MuSiCa s'ha establert com una de les aplicacions web de referència per al càlcul de la càrrega mutacional tumoral i la caracterització de les signatures mutacionals segons les signatures de referència de COSMIC, sent àmpliament utilitzada des de la seva publicació.

3. La classificació de mostres per *clustering* i anàlisi de components principals segons les contribucions de les diferents signatures mutacionals és una característica distintiva de MuSiCa, que no està disponible en cap de les aplicacions competidores que existeixen per realitzar anàlisis de signatures mutacionals.

4. La caracterització molecular de mostres somàtiques de CCR procedents del projecte TCGA es va replicar de forma senzilla i precisa a través de l'anàlisi de signatures mutacionals de MuSiCa.

5. L'anàlisi integrada de dades de SEC germinals i tumorals, tenint en compte diferents classes de variants genètiques i basat en la hipòtesi clàssica dels dos *hits* de Knudson i la caracterització mutacional somàtica, s'ha demostrat útil per a la identificació de nous GSTs candidats a estar involucrats en la predisposició al CCR familiar.

6. Es van identificar sis gens com a potencials candidats per a la predisposició germinal al CCR familiar, incloent gens coneguts per la seva implicació en la

predisposició a altres neoplàsies, com és el cas de *BLM*, *BRCA2* i *ERCC2*, així com gens associats a la reparació de l'ADN , *RECQL*, *REV3L* i *RIF1*.

7. L'anàlisi del perfil mutacional somàtic pot ser útil en el descobriment del defecte germinal responsable. En el nostre estudi, això es va exemplificar amb un gen candidat lligat a la reparació de l'ADN, *RECQL*, que es va trobar mutat en l'ADN germinal d'un pacient amb un fenotip hipermutat en el tumor, reforçant el paper potencial d'aquest gen en el CCR hereditari.

Summary in Spanish | Resumen en castellano

Identificación de nuevos genes candidatos para la predisposición germinal al cáncer colorrectal familiar a través de la caracterización mutacional somática

Introducción

El cáncer colorrectal (CCR) es una de las neoplasias malignas más comunes y con mayor mortalidad asociada en el mundo, con más de un millón y medio de nuevos casos y más de 800.000 muertes cada año (**Figura 1**) (Bray et al., 2018). La mayor incidencia se encuentra en las regiones más desarrolladas, incluyendo Australia, Nueva Zelanda, Europa, Asia oriental y Norteamérica (**Figura 2**). En Europa, el CCR representa el segundo tipo de cáncer en incidencia y mortalidad considerando ambos sexos, mientras que en España se trata del primero en incidencia y sólo está detrás del cáncer de pulmón en mortalidad (Ferlay et al., 2019). Como enfermedad compleja, la etiología del CCR implica la combinación de diferentes factores de riesgo. Además de factores no modificables, como la edad o el sexo masculino, los factores ambientales han sido asociados con un aumento en la incidencia de CCR, particularmente con la denominada occidentalización de la dieta y el estilo de vida (Brenner et al., 2014).

El CCR fue uno de los primeros tumores sólidos caracterizados a nivel molecular, con distintas vías de señalización implicadas en el inicio y la progresión de la carcinogénesis (Fearon, 2011). Este proceso se describió inicialmente a través de la secuencia adenoma-carcinoma, donde una acumulación de alteraciones genéticas en oncogenes y genes supresores de tumores (GSTs) da lugar a una transición de una lesión precursora (llamada pólipo o adenoma) a un carcinoma, a través de diferentes estados intermedios caracterizados por alteraciones genéticas y/o epigenéticas específicas (**Figura 3**) (Vogelstein et al., 1988; Kuipers et al., 2015). Los oncogenes se definen como aquellos genes cuya activación acelera el desarrollo tumoral, mientras que en los GSTs, al contrario, es su pérdida de expresión la que está ligada a la adquisición del fenotipo neoplásico (Bashyam et al., 2019). Este fenotipo se caracteriza principalmente por un crecimiento celular descontrolado y la supresión de los mecanismos de muerte y reparación celulares, así como por la adquisición de las capacidades de invasión y metástasis (**Figura 4**) (Hanahan & Weinberg, 2000, 2011). El defecto molecular inicial en la mayoría de tumores colorrectales (más de un 70%) sucede en el GST *APC*, causando la desregulación de la vía de señalización Wnt/β-catenina (Kinzler & Vogelstein, 1996; Brenner et al., 2014), aunque otras vías de señalización se ven también afectadas durante la transformación neoplásica, incluyendo RAS–RAF–MAPK, PI3K–AKT, TGFβ y

p53 (Kuipers et al., 2015). Recientemente se ha identificado una vía de carcinogénesis colorrectal alternativa, iniciada por una tipología de lesiones precancerosas diferenciada, las lesiones serradas, que actualmente se conoce que representan más del 15% de los casos de CCR y que presentan características histológicas y moleculares diferenciadas respecto a los adenomas convencionales (**Figura 3**) (Carballal et al., 2013; IJspeert et al., 2015).

A nivel molecular se consideran tres vías principales para la carcinogénesis colorrectal: inestabilidad cromosómica (INC), inestabilidad de microsatélites (IMS) y la caracterizada por un fenotipo de hipermetilación de islas de dinucleótidos CpG (CIMP, por sus siglas en inglés *CpG island methylator phenotype*) (**Figura 5**). La INC, caracterizada por la acumulación de alteraciones del número de copia, fue la primera vía molecular descrita y se conoce que es origen de la mayor parte de casos de CCR, especialmente de los casos esporádicos (hasta un 85% de estos últimos). Respecto a la IMS, se define por alteraciones en los microsatélites (secuencias repetitivas de ADN localizadas a lo largo del genoma), que aparecen en forma de pequeñas inserciones o deleciones (indels), dando lugar a mutaciones de terminación de la proteína por cambio en la pauta de lectura. Estas mutaciones deberían ser corregidas por el sistema de reparación del ADN denominado reparación de mal apareamiento de bases (MMR, del inglés *mismatch repair*). Cuando este sistema no funciona correctamente, aparece el fenotipo de IMS, ampliamente utilizado como biomarcador para la detección de un MMR deficiente en CCR y ligado a hipermutación. Por su parte, el CIMP se liga a la hipermetilación de los promotores de numerosos GSTs asociados al cáncer, lo que provoca la supresión de su transcripción (Carethers & Jung, 2015; Kuipers et al., 2015). Recientemente se ha descrito una nueva clasificación molecular para el CCR basada en patrones de expresión génica, los denominados subtipos moleculares consenso (**Figura 6**) (Guinney et al., 2015; Dienstmann et al., 2017).

La predisposición germinal a enfermedades complejas, como es el caso del CCR, implica una distribución diversa de variantes genéticas, que pueden ser clasificadas según su frecuencia en la población, así como respecto a su riesgo asociado a desarrollar una determinada enfermedad (conocido como penetrancia) (**Figura 7**) (McCarthy et al., 2008; Manolio et al., 2009). Las variantes de alta penetrancia se definen como aquellas que causan un mayor efecto en la susceptibilidad a la enfermedad, pero que comúnmente son más raras en la población. Se han ligado a enfermedades que siguen un patrón mendeliano (Mendel, 1866), donde la alteración de un único gen es frecuentemente la responsable del fenotipo. Estas variantes se han identificado clásicamente a través de estudios de ligamiento, también en el caso de los síndromes hereditarios de predisposición al CCR (**Figura 8**) (Bodmer et al., 1987; Lindblom et al., 1993; Peltomaki et al., 1993). Por otro lado, las variantes de baja penetrancia se caracterizan por ser comunes en la población general y tener un pequeño efecto

individualmente en el desarrollo de la enfermedad. Sin embargo, una combinación de estas variantes, junto con la interacción con factores de riesgo ambientales puede contribuir significativamente a la predisposición a la enfermedad. Se han detectado mayoritariamente por estudios de asociación del genoma completo (GWAS, del inglés *genome wide association studies*), que en caso del CCR han permitido identificar alrededor de 130 variantes implicadas y que explican un 7-8% de la susceptibilidad asociada a esta enfermedad (**Figura 8**) (Jiao et al., 2014; Peters et al., 2015; Buniello et al., 2019). En determinadas enfermedades, como es el caso del CCR, la ratio de heredabilidad estimada respecto a los estudios clásicos en gemelos y familias (12-35%) no está de acuerdo con la heredabilidad explicada por las variantes genéticas con una asociación conocida con la enfermedad (2-8%), por lo que esto conlleva una heredabilidad *no filiada* (Jiao et al., 2014; Valle, Vilar, et al., 2019). Esta heredabilidad estaría relacionada en parte con aquellas variantes no suficientemente frecuentes para ser identificadas por GWAS pero tampoco con un efecto en el desarrollo de la enfermedad suficiente para ser detectadas por estudios familiares de ligamiento (**Figuras 7-8**) (Manolio et al., 2009). En este sentido, la secuenciación de nueva generación (SNG) se ha desmarcado como la herramienta más utilizada para la identificación de estas variantes. Esta técnica ha revolucionado el campo de la genética, permitiendo la identificación de distintas clases de variantes implicadas en la predisposición a distintas enfermedades a un bajo coste relativo, incluyendo principalmente variantes de un único nucleótido (SNVs, del inglés *single nucleotide variants*) e indels, pero también variantes de número de copia (CNVs, del inglés *copy number variants*) (Lappalainen et al., 2019). Las CNVs se definen como fragmentos de ADN de un tamaño superior a 50 nucleótidos con variaciones en el número de copia (deleciones o duplicaciones) respecto al genoma de referencia (Alkan et al., 2011). La aplicación de la SNG más exitosa en los estudios biomédicos traslacionales ha sido la secuenciación del exoma completo (SEC), es decir, de todas las regiones codificantes del genoma (Teer & Mullikin, 2010). Sin embargo, de cara a la identificación de nuevos genes de predisposición, esta tecnología necesita la implementación de una estrategia de priorización, que posibilite reducir el alto número de variantes que se identifican inicialmente (**Figura 9**) (Ott et al., 2015).

Los síndromes hereditarios de predisposición al CCR relacionados con variantes genéticas de alta penetrancia representan el 2-8% de todos los casos, y hasta el 6-10% si se consideran también las variantes de penetrancia moderada. Distintos genes, pertenecientes a diferentes vías de señalización, han sido implicados en estos síndromes, caracterizados por estar originados por diferentes tipologías de lesiones preneoplásicas (o pólipos) (**Figura 10**) (Tomlinson, 2015). Se clasifican fenotípicamente según la presencia o no de una acumulación de estas lesiones precursoras denominada poliposis (**Figura 11**) (Valle, Vilar, et al., 2019). Los síndromes polipósicos se dividen a su

vez según el tipo de pólipos encontrados en los pacientes. Con poliposis adenomatosa están la poliposis adenomatosa familiar y su variante atenuada (ligadas fundamentalmente a mutaciones germinales en el gen *APC*) (Leppert et al., 1987, 1990), la poliposis asociada a *MUTYH* (Al-Tassan et al., 2002), la poliposis asociada a la reparación del ADN por corrección realizada por las polimerasas (ligada a defectos germinales en *POLE* y *POLD1)* (Palles et al., 2013) y el síndrome tumoral asociado a *NTHL1* (implicado en hasta 14 tipos tumorales diferentes) (Weren, Ligtenberg, et al., 2015). Por su parte, originado por pólipos serrados, aparece el síndrome de poliposis serrada (del que sólo se ha propuesto un gen candidato para su predisposición hereditaria, *RNF43*, aunque con controversia) (Gala et al., 2014); surgiendo de pólipos hamartomatosos, el síndrome de Peutz-Jeghers (ligado a defectos germinales en *STK11*) (Giardiello et al., 1987), el síndrome de poliposis juvenil (*BMPR1A*, *SMAD4*) (Howe et al., 1998, 2001) y el síndrome tumoral *PTEN*-hamartoma / síndrome de Cowden (*PTEN*) (Liaw et al., 1997); y a través de una combinación de las tres tipologías de pólipos, el síndrome hereditario de poliposis mixta (*GREM1*) (Jaeger et al., 2012). Por otro lado, respecto a los síndromes no polipósicos, destaca el síndrome de Lynch. Éste está asociado a mutaciones germinales en los genes del sistema de MMR (*MLH1*, *MSH2*, *MSH6*, *PMS2*) y constituye el síndrome de CCR hereditario más frecuente (H. T. Lynch et al., 2015). Debido a esto último, se han desarrollado una serie de guías clínicas para la identificación de las familias con más probabilidad de ser portadores de este síndrome (**Figura 12**) (Vasen et al., 1999; Umar et al., 2004). Los mencionados síndromes presentan en general una herencia autosómica dominante, excepto en el caso de aquellos ligados a mutaciones en genes de la vía de reparación del ADN por escisión de bases (BER, del inglés *base excision repair*), *MUTYH* y *NTHL1*, cuyo patrón de herencia es autosómico recesivo (Valle, Vilar, et al., 2019).

Además de los comentados síndromes de predisposición hereditarios (que explican hasta un 8% de la heredabilidad), se especula con que los factores genéticos estén detrás de un 12-35% del total de casos de CCR (Lichtenstein et al., 2000; Jiao et al., 2014; Peters et al., 2015). Esta heredabilidad *no filiada* ha sido objeto de estudio en los últimos años con el objetivo de la identificación de nuevos genes candidatos, que podrían tener un fuerte impacto de cara al consejo genético en las familias afectadas. La SNG ha sido la tecnología principalmente utilizada en este esfuerzo de identificación de nuevos genes implicados en la predisposición al CCR (Valle, de Voer, et al., 2019). De este modo, un gran número de genes candidatos ha sido propuesto por diferentes grupos de investigación, incluyendo *BUB1*, *BUB3* (de Voer et al., 2013), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *BLM* (de Voer et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *MIA3* (Schubert et al., 2017), *SETD6* (Martín-Morales et al., 2017) y *BRF1* (Bellido et al., 2018) como los más prometedores de acuerdo a los estudios funcionales realizados y a la validación en cohortes de CCR familiar adicionales.

Según la hipótesis de los dos *hits* de Knudson, el desarrollo neoplásico comienza con dos eventos mutacionales en un único gen (un GST), que impiden su expresión. Así, las diferencias observadas entre las formas hereditarias y esporádicas/no hereditarias de un determinado cáncer se deben a la diferente combinación de estas alteraciones genéticas o *hits* (que pueden ser de diferentes clases: SNVs, indels, CNVs, pérdidas de heterocigosidad (LOH, del inglés *loss of heterozygosity*) o alteraciones en la metilación). En el caso de un cáncer hereditario existiría una primera alteración en el ADN germinal seguida por un segundo *hit* somático, mientras que en los casos esporádicos se encontrarían directamente dos mutaciones en las células tumorales (**Figura 13**). Así, se explicaría la aparición más temprana de los cánceres hereditarios, ya que sólo es necesario un evento mutacional en el tumor para el desarrollo de la enfermedad (Knudson, 1971).

Todos los cánceres se caracterizan por múltiples mutaciones somáticas. Asimismo, éstas se clasifican en mutaciones *driver* o *passenger* según sus efectos en el desarrollo tumoral (Stratton et al., 2009). Aunque la identificación de mutaciones *driver* se ha priorizado en la mayoría de los estudios de secuenciación, al ser las que se seleccionan positivamente y están detrás del desarrollo carcinogénico, las mutaciones *passenger* también han demostrado ser informativas. De hecho, el número total de mutaciones acumuladas por un tumor (denominado como carga mutacional tumoral (TMB, del inglés *tumor mutational burden*)), altamente variable entre tipos tumorales y también dentro del mismo cáncer (**Figura 14**) (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013), ha emergido en los últimos años como un prometedor biomarcador para inmunoterapias, debido a su relación con la carga de neoantígenos (Chalmers et al., 2017).

Además de la caracterización de la TMB, las mutaciones *passenger* también son responsables de la aparición de un nuevo campo de estudio en los últimos años. Asumiendo que los patrones de estas mutaciones no varían en el tiempo, pueden ser utilizadas como una imagen representativa de los mecanismos mutacionales que han permanecido activos durante el proceso carcinogénico (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Cada proceso mutacional deja una huella particular en el genoma de una célula, un perfil de mutaciones específico denominado firma mutacional. Mecanismos celulares endógenos, como la replicación y la reparación del ADN, pueden generar mutaciones debido a su tasa de error intrínseca. Por otro lado, las mutaciones también pueden ser debidas a exposiciones mutagénicas exógenas, como sería el caso del tabaco o la luz ultravioleta. Así, el conjunto final de mutaciones recogido en un tumor está determinado por la intensidad y la duración de todos los procesos mutacionales activos durante el desarrollo neoplásico (**Figura 15**) (Nik-Zainal et al., 2012; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013).

El daño al ADN puede aparecer en forma de diferentes tipos de variantes genéticas, aunque para la descripción de las firmas mutacionales se han utilizado hasta el momento principalmente las SNVs por razones técnicas. Así, en el conjunto actual de firmas mutacionales de referencia se consideran seis tipos de cambio de nucleótido, según la pirimidina mutada de la pareja de bases de Watson-Crick, incluyendo cuatro posibles transversiones, C>A, C>G, T>A y T>G, y dos transiciones, C>T y T>C. Para una caracterización más estricta de los procesos mutacionales responsables de las mutaciones, se tienen en cuenta también las bases adyacentes al cambio en los contextos 5' y 3', dando lugar a un total de 96 posibilidades (6 sustituciones de bases * 4 nucleótidos anteriores * 4 nucleótidos posteriores) (**Figura 16**). De este modo, cada firma mutacional se compone por una distribución única de estos 96 posibles tipos de mutaciones (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013). En los últimos años se ha generado un modelo matemático que ha permitido la detección y cuantificación precisa de cada una de las firmas mutacionales asociadas a los distintos procesos mutagénicos implicados en el cáncer. Para ello, se utilizó inicialmente un algoritmo basado en la factorización matricial no negativa denominado SigProfiler, que fue implementado utilizando MATLAB (Alexandrov, Nik-Zainal, Wedge, Campbell, et al., 2013). Recientemente, este algoritmo ha sido traducido a otros lenguajes de programación abiertos (Gehring et al., 2015; Blokzijl et al., 2018), mientras que también han surgido nuevas estrategias computacionales para la identificación de firmas mutacionales (Kasar et al., 2015; Shiraishi et al., 2015; Baez-Ortega & Gori, 2019).

A través de estos modelos computacionales ha sido posible la extracción de firmas mutacionales de referencia, cada una de ellas asociada a un proceso mutagénico específico del cual, en algunos casos, se ha podido identificar su etiología. Estas firmas de referencia permiten que el análisis de firmas mutacionales no sólo se restrinja a la identificación agnóstica de nuevas firmas (conocido como análisis *de novo*, es decir, sin utilizar ningún conocimiento previo). También hacen posible la caracterización de los procesos mutacionales implicados a nivel de muestra respecto a una referencia (denominada como análisis de ajuste de firmas mutacionales). Para llevar a cabo esta tipología de análisis, más orientada a su aplicación en la práctica clínica, también han surgido nuevas herramientas bioinformáticas en los últimos años (Rosenthal et al., 2016; Blokzijl et al., 2018). Sin embargo, todavía están orientadas a expertos bioinformáticos, permaneciendo inaccesibles para una parte importante de la comunidad científica. El número de firmas de referencia ha ido creciendo paulatinamente, a medida que el número de muestras tumorales analizadas se ha ido incrementando, lo que se debe al sucesivo aumento de potencia estadística del modelo matemático. En una primera aplicación de esta metodología se extrajeron cinco firmas mutacionales de SNVs (posteriormente reducidas a cuatro después de una optimización del modelo) de una cohorte de 21 muestras de cáncer de mama (Nik-Zainal et al., 2012; Alexandrov, Nik-

Zainal, Wedge, Campbell, et al., 2013). A continuación, el conjunto de referencia de firmas mutacionales se amplió a 21 (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al., 2013), y posteriormente a 30 después de la aplicación de esta metodología a unas 12.000 muestras de 40 tipologías diferentes de cáncer (**Figura 17**) (Alexandrov et al., 2015; Tate et al., 2018). Este conjunto de 30 firmas mutacionales de referencia se ha utilizado en la gran mayoría de publicaciones que han realizado análisis de firmas mutacionales hasta la fecha (Grolleman, Díaz-Gay, et al., 2019), y se puede encontrar como parte de la base de datos COSMIC (versión 2 – marzo 2015) (Wellcome Trust Sanger Institute, 2019a). Finalmente, el actual conjunto de firmas mutacionales de referencia se ha extraído de más de 23.000 muestras de cáncer y se compone de 49 firmas de SNVs (también denominadas firmas de SBSs, del inglés *single base substitutions*), mientras que ha incorporado también otras tipologías de variantes, incluyendo 17 firmas asociadas a indels y 11 ligadas a sustituciones de dos bases consecutivas (**Figura 18**) (Alexandrov et al., 2019). También se encuentra disponible en COSMIC (versión 3 – mayo 2019) (Wellcome Trust Sanger Institute, 2019b).

Se ha encontrado una distribución diferente de firmas mutacionales en los diferentes tejidos, lo que está de acuerdo con las distintas ratios de reemplazamiento celular, así como con la distinta influencia de las exposiciones ambientales según el tejido en cuestión. Algunas firmas, como es el caso de SBS1, SBS5 y SBS40, se han asociado con la edad de diagnóstico, por tanto reflejando la influencia del proceso de envejecimiento en la carcinogénesis (Alexandrov et al., 2015, 2019). Respecto al CCR, según la información disponible en COSMIC y una serie de publicaciones recientes, existe una contribución de diferentes firmas mutacionales, incluyendo las mencionadas firmas relacionadas con el envejecimiento. Sin embargo, estas firmas contribuyen a un número reducido de mutaciones, en comparación con aquellas relacionadas específicamente con dos conocidos defectos moleculares presentes en el CCR: las deficiencias en los procesos de reparación del ADN por MMR y por corrección de las polimerasas (7 firmas diferenciadas en el caso de un mal funcionamiento del sistema de MMR: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 y SBS44, y las firmas SBS10a y SBS10b en el caso de mutaciones en el dominio exonucleasa de la polimerasa *POLE*) (Nagahashi et al., 2016; Alexandrov et al., 2019). Adicionalmente, algunas de estas firmas se han asociado a la concurrencia de alteraciones genéticas en ambos sistemas de reparación del ADN, incluyendo las firmas SBS14 (mutación en *POLE* y en los genes del MMR) y SBS20 (mutación en *POLD1* y MMR defectuoso), siendo encontrada también esta última en casos de CCR (Haradhvala et al., 2018; Alexandrov et al., 2019). Recientemente se ha ampliado el espectro de firmas mutacionales asociadas al CCR, con la inclusión de dos firmas relacionadas con el sistema BER de reparación del ADN y específicamente con defectos en dos conocidos genes de predisposición al CCR: *MUTYH* (firma SBS36) (Pilati et al., 2017; Viel et al., 2017) y *NTHL1* (SBS30) (Drost et al., 2017;

Grolleman, de Voer, et al., 2019). Otras firmas también han sido asociadas al CCR, aunque con un rol menos prevalente, como es el caso de SBS2, SBS13 (ligadas a la actividad de las deaminasas APOBEC), SBS3 (sistema de reparación del ADN por recombinación homóloga defectuoso y mutaciones en *BRCA1/2*), SBS9 (actividad de la polimerasa eta), SBS17a, SBS17b, SBS18 (daño al ADN provocado por las especies reactivas del oxígeno), SBS12, SBS28, SBS37 y SBS41 (etiología desconocida), así como una nueva firma identificada por Roerink y colaboradores en un estudio reciente (Nagahashi et al., 2016; Roerink et al., 2018; Alexandrov et al., 2019).

Las firmas mutacionales, así como la TMB, pueden ser utilizadas para la identificación de los defectos genéticos germinales que han estado activos durante el origen y la evolución de un determinado cáncer. Esto es particularmente evidente para aquellas firmas mutacionales con una etiología conocida y, en particular, para aquellas asociadas a procesos mutacionales responsables de síndromes hereditarios de predisposición al cáncer, como los derivados de defectos en los mecanismos de reparación del ADN (**Figura 19**) (J. Ma et al., 2018; Van Hoeck et al., 2019). En el caso de los sistemas de corrección de las polimerasas y MMR, los defectos germinales se han identificado ligados a una TMB alta, es decir, a tumores hipermutados, y adicionalmente a las mencionadas firmas mutacionales características (Muzny et al., 2012; Kandoth et al., 2013; Alexandrov et al., 2019). Las firmas mutacionales, así como el análisis de la TMB, podrían ayudar en la identificación y el descubrimiento de los procesos mutacionales responsables de los diferentes síndromes hereditarios de cáncer, así como favorecer el diagnóstico genético y la selección de tratamientos en los pacientes, como se ha demostrado recientemente en el caso de las alteraciones en los genes *BRCA1/2* y *NTHL1* (Davies et al., 2017; Grolleman, de Voer, et al., 2019).

Hipótesis

El CCR es una enfermedad compleja y, por tanto, con una etiología en la que se entremezclan factores genéticos y ambientales. La predisposición genética está detrás de hasta un 35% de los CCRs según estudios familiares y de gemelos, mientras que los síndromes de predisposición conocidos y asociados a defectos genéticos germinales específicos sólo explican un 2-8% de los casos. De esta forma, se observa una heredabilidad *no filiada* para esta neoplasia. La SNG es la técnica más adecuada para llevar a cabo la identificación de nuevos genes implicados en la predisposición al CCR, como se ha demostrado en estudios recientes en genes como *POLD1*, *POLE* y *NTHL1*. Sin embargo, esta tecnología identifica un gran número de variantes genéticas en cada paciente, generando así la necesidad de una estrategia de priorización. En este sentido, según la clásica hipótesis de los dos *hits* de Knudson, además de las alteraciones genéticas germinales, también las somáticas pueden jugar un rol fundamental en proporcionar nuevo conocimiento respecto a la predisposición hereditaria al CCR. De

acuerdo con esto, el análisis del perfil mutacional somático se ha utilizado recientemente para la identificación de nuevos genes de predisposición al CCR, así como un biomarcador prometedor de cara al diagnóstico, pronóstico y tratamiento de esta neoplasia. Aunque se han desarrollado diversos paquetes bioinformáticos para realizar este tipo de análisis, todavía permanece inaccesible para una proporción sustancial de la comunidad científica.

Objetivos

El objetivo principal de la presente tesis doctoral es el de identificar nuevos genes candidatos que puedan estar implicados en la predisposición germinal al CCR familiar. Un análisis combinado germinal-tumoral de datos de SEC y una aplicación bioinformática para realizar caracterización mutacional somática se desarrollarán para ser utilizadas como estrategias de priorización.

Con este fin, se llevarán a cabo los siguientes objetivos específicos:

1. Desarrollo de una aplicación computacional para realizar análisis de los perfiles mutacionales somáticos, a través de una interfaz sencilla adecuada para investigadores no especializados en bioinformática y accesible libremente mediante una página web. Estarán disponibles tanto la caracterización de la TMB como el ajuste de las firmas mutacionales según las firmas de referencia versión 2 de COSMIC, así como la clasificación de muestras por *clustering* y análisis de componentes principales.

2. Análisis integrado basado en la hipótesis de los dos *hits* de Knudson de datos de SEC procedentes de ADN germinal y tumoral de una cohorte de 18 pacientes de CCR familiar, con el objetivo de identificar nuevos GSTs potenciales. Se tendrán en cuenta distintas clases de alteraciones genéticas, mientras que los genes candidatos se seleccionarán cuando tanto el ADN germinal como el tumoral estén afectados por una de estas alteraciones.

3. Caracterización somática mutacional de la mencionada cohorte de CCR familiar utilizando la herramienta bioinformática desarrollada previamente, a través del análisis de la carga mutacional tumoral y las firmas mutacionales.

Resultados y discusión

El primero de los estudios publicados como parte de esta tesis doctoral presenta el desarrollo de la aplicación MuSiCa (del inglés *Mutational Signatures in Cancer*), que constituye una de las primeras herramientas web disponibles para realizar una caracterización mutacional somática completa de los tumores secuenciados con técnicas de SNG.

Tanto el cálculo de la TMB como la reconstrucción de los perfiles mutacionales somáticos según las firmas mutacionales de referencia versión 2 de COSMIC (Wellcome

Trust Sanger Institute, 2019a) están disponibles en MuSiCa. Una futura actualización de la aplicación será necesaria de cara a la adaptación a la nueva versión 3 de estas firmas de referencia, que deberá incluir las nuevas clases de variantes a tener en cuenta. Respecto al ajuste de firmas mutacionales, MuSiCa utiliza como base el paquete de R/Bioconductor MutationalPatterns (Blokzijl et al., 2018), que se basa en la resolución de un problema de optimización de mínimos cuadrados no negativos a través de un algoritmo de método de conjunto activo (Lawson & Hanson, 1974) incluido en el paquete de R pracma (Borchers, 2019). MuSiCa proporciona una interfaz gráfica a este paquete, creada a través del paquete de R Shiny (W. Chang et al., 2019) y específicamente diseñada para investigadores no especializados en bioinformática, así como algunas características adicionales. MuSiCa está disponible de forma gratuita como parte de la página web de nuestro grupo de investigación (http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html), lo que permite su uso de forma sencilla por parte de cualquier miembro de la comunidad científica sin necesidad de grandes recursos a nivel de computación. De hecho, según los datos recogidos por la plataforma Google Analytics durante los primeros 14 meses desde la publicación del artículo de MuSiCa, 1.344 usuarios únicos de un total de 53 países diferentes han accedido a la web de la aplicación, en un total de 3.045 sesiones (**Figura 20**). También es posible utilizar MuSiCa de forma local, para lo cual las dependencias requeridas para su instalación, así como el código fuente en R, están libremente disponibles en GitHub (https://github.com/marcos-diazg/musica).

MuSiCa permite una caracterización del perfil mutational somático a nivel de muestra, lo que proporciona grandes beneficios en el caso de pequeñas cohortes y muestras individuales (Blokzijl et al., 2018). Ambos escenarios son comunes en el entorno clínico, donde el perfil mutacional de cada paciente debería ser contrastado respecto al mismo conjunto de firmas mutacionales de referencia (Rosenthal et al., 2016; Baez-Ortega & Gori, 2019). Así, MuSica se establece como una herramienta útil para la caracterización de firmas mutacionales en la práctica clínica, siempre que se disponga de datos de SNG tanto de ADN germinal como tumoral (ya que es necesario para poder identificar las variantes somáticas).

Otras aplicaciones web han sido desarrolladas en los últimos años para realizar análisis de firmas mutacionales (Baez-Ortega & Gori, 2019; Grolleman, Díaz-Gay, et al., 2019; Hanane et al., 2019). Pmsignature fue la primera herramienta web que dispuso de una interfaz gráfica, aunque sólo permitía el descubrimiento de firmas mutacionales *de novo* (a través de su novedoso modelo probabilístico) y no el ajuste según un conjunto de firmas de referencia (Shiraishi et al., 2015). Por su parte, la aplicación web MutaGene proporciona un marco computacional para una completa caracterización de las mutaciones tumorales y los procesos mutacionales asociados, permitiendo analizar genes específicos y buscar potenciales mutaciones *driver*. Aunque está enfocada en la

evaluación de datos de muestras de cáncer disponibles públicamente, también permite realizar análisis de ajuste de firmas mutacionales, pero en este caso sólo permite analizar las muestras de una en una, limitando así la comparación en cohortes de más de un paciente (Goncearenco et al., 2017). Al igual que en el caso anterior, mSignatureDB proporciona la posibilidad tanto de analizar datos de muestras tumorales disponibles públicamente como de realizar un análisis de firmas mutacionales en una serie de muestras proporcionadas directamente por los usuarios. Este análisis puede ser *de novo* (utilizando el paquete mutSignatures (Fantini et al., 2018)) o mediante ajuste de firmas (a través de deconstructSigs (Rosenthal et al., 2016)), el cual presenta un tiempo de computación muy superior al de MutationalPatterns y, por tanto, de MuSiCa (P.-J. Huang et al., 2018). Por último, Mutalisk constituye la aplicación web más completa respecto al análisis mutacional somático a nivel de muestra hasta la fecha. Además de la descomposición de firmas, Mutalisk proporciona información sobre hipermutación localizada (denominada *kataegis*), sesgo de la cadena transcripcional, contenido de GCs, tiempo de replicación del ADN, modificaciones de las histonas e hipersensibilidad a la DNasa I (J. Lee et al., 2018). Respecto a sus competidoras, MuSiCa presenta funcionalidades exclusivas de cara a la clasificación de muestras, que se puede realizar a través de *clustering* y análisis de componentes principales, y que podría tener un importante potencial en el entorno clínico. Así, por ejemplo, en una cohorte de un cierto subtipo de cáncer muy específico y con un fenotipo bien definido, la comparación de sus perfiles de firmas mutacionales con otros de pacientes de otras tipologías de cáncer podría proporcionar nuevo conocimiento respecto al defecto genético responsable. Esta estrategia ha sido utilizada satisfactoriamente en el caso de la deficiencia de *NTHL1* y su asociación con la firma SBS30 (Grolleman, de Voer, et al., 2019).

Como medida de la potencial aplicabilidad de MuSiCa se llevó a cabo la replicación de la caracterización de los perfiles mutacionales somáticos de los tumores de colon procedentes del proyecto TCGA (Muzny et al., 2012). Se utilizaron un total de 433 muestras y se consiguió reproducir satisfactoriamente las vías moleculares de IMS (dominada por las firmas asociadas a un MMR defectuoso: SBS6, SBS15, SBS20 y SBS26), deficiencia en el sistema de reparación por corrección de las polimerasas (ligada a la firma asociada a mutaciones en *POLE: SBS10*) e INC (que se encontró dominada por la firma asociada a la edad SBS1). Esto último se debe a que las alteraciones *driver* en esta vía son principalmente de número de copia, mientras que el análisis de firmas únicamente considera las SNVs, que en este caso serían eventos *passenger* ligados al proceso de envejecimiento.

Por otro lado, en el segundo estudio de esta tesis doctoral se ha desarrollado y aplicado un análisis integrado de datos de SEC germinal y tumoral en una cohorte de 18 pacientes no relacionados de CCR familiar, junto con una caracterización de los perfiles

mutacionales somáticos realizada con la aplicación MuSiCa desarrollada previamente, con el objetivo de encontrar nuevos genes candidatos responsables de la predisposición germinal a esta neoplasia.

Las muestras utilizadas en este estudio pertenecen a una cohorte más amplia de CCR familiar (71 pacientes de 38 familias), de la que se dispone de datos de SEC germinal y que ha sido utilizada previamente en diversos estudios del grupo de investigación (Esteban-Jurado et al., 2015, 2016; Franch-Expósito et al., 2018). Estas familias fueron seleccionadas por tener una fuerte agregación para la enfermedad, así como por no presentar defectos germinales en los genes de predisposición ya conocidos. La posibilidad de disponer de datos de secuenciación combinados germinales y tumorales proporcionó, por primera vez en nuestro grupo de investigación, la oportunidad de analizar el perfil de alteraciones genéticas somáticas. En este sentido, se ha explotado y trasladado al ámbito somático la experiencia acumulada en la identificación y el análisis de distintas tipologías de variantes potencialmente patogénicas en datos de SEC germinal (incluyendo SNVs, indels y CNVs).

Después de la identificación de variantes a través de diferentes softwares (GATK HaplotypeCaller para SNVs/indels germinales, CoNIFER y ExomeDepth para CNVs germinales, MuTect2 para SNVs/indels somáticas y ALFRED para predecir LOHs somáticas), se utilizó un análisis integrado germinal-tumoral basado en la hipótesis de los dos *hits* de Knudson para la priorización de los genes más interesantes como candidatos a la predisposición al CCR. Así, estos GSTs candidatos debían presentar una alteración germinal y otra somática de forma que se perdiese completamente su función. Esta estrategia ha sido también usada en algunos estudios recientes. En el caso de Spier y colaboradores, utilizaron esta estrategia en una cohorte de 7 pacientes de poliposis adenomatosa, aunque no pudieron identificar ningún gen candidato que siguiese el modelo de los dos *hits* (Spier et al., 2016). Por otro lado, en un análisis de más de 10.000 muestras de distintos tipos de cáncer públicamente accesibles, se detectaron un total de 13 genes, incluyendo genes de predisposición a diferentes neoplasias ya conocidos, como *BRCA1*, *BRCA2* y *ATM*, pero también nuevos potenciales candidatos como es el caso de la histona metiltransferasa *NSD1* (Park et al., 2018).

A pesar de que el análisis realizado en nuestro estudio tiene en cuenta distintos tipos de variantes, otras posibles alteraciones podrían actuar también como primer o segundo *hit* en el modelo de Knudson, incluyendo alteraciones epigenéticas, como modificaciones de las histonas o ARNs no codificantes (microARNs o ARNs no codificantes largos) (Okugawa et al., 2015), así como defectos en regiones no codificantes del genoma (que no han podido ser evaluadas por ser los datos de SNG de partida procedentes de SEC). Por otro lado, la estrategia de priorización escogida limita a su vez la selección de candidatos, ya que mecanismos como la haploinsuficiencia

hacen prescindible el segundo *hit* somático para que el gen afectado a nivel germinal tenga una influencia en la predisposición hereditaria (Deutschbauer et al., 2005), como se ha visto en el caso de los genes *BUB1* y *BUB3* (de Voer et al., 2013). Además, se podrían haber utilizado otras estrategias de priorización, como la replicación en cohortes adicionales o los estudios funcionales, que han resultado satisfactorias en los casos de genes candidatos como *RPS20* (Nieminen et al., 2014), *SEMA4A* (Schulz et al., 2014), *FAN1* (Seguí et al., 2015), *FOCAD* (Weren, Venkatachalam, et al., 2015), *SETD6* (Martín-Morales et al., 2017) o *BRF1* (Bellido et al., 2018).

Después de la realización de un control de calidad, que descartó dos de las 18 muestras por una baja calidad de secuenciación tumoral, se aplicaron las correspondientes *pipelines* de análisis germinal y somático. En la cohorte final de 16 muestras, se encontraron 494 SNVs y 42 indels germinales, mientras que ninguna CNV germinal fue identificada afectando a genes con una función compatible con la predisposición al CCR familiar. Considerando las alteraciones somáticas, se detectó un total de 143 genes con variantes tanto en el genoma germinal como en el somático (**Figura 21**). En tres de estos genes, *ADCY8*, *HSPG2* y *TTN*, se identificaron dos SNVs, una germinal y una somática, aunque el gen *TTN* fue descartado por su gran longitud (que podría causar la acumulación de variantes simplemente por azar) (Chauveau et al., 2014). Por otro lado, en 141 genes se identificó una SNV o indel germinal y se predijo un LOH tumoral como segundo *hit* (también incluyendo al mencionado *HSPG2*). Para reducir el número de genes a una primera selección de 16 candidatos potenciales, fue necesario un proceso de curado manual según la información funcional previamente publicada para cada gen (**Figura 21**). Cabe destacar que se encontró un enriquecimiento de la reparación del ADN entre las funciones asociadas a los genes seleccionados (7 de 16 genes implicados, incluyendo *BLM*, *BRCA2*, *ERCC2*, *PARP2*, *RECQL*, *REV3L* y *RIF1*), lo que está de acuerdo con parte de los genes hereditarios clásicos de CCR (Valle, Vilar, et al., 2019). También se destacaron aquellos genes causantes de un síndrome de predisposición a cáncer cuando se encuentran mutados germinalmente (*BLM*, *BRCA2*, *ERCC2* y *SMARCA4*) (Rahman, 2014), así como dos genes asociados con síndromes de predisposición al CCR conocidos, el síndrome de Cowden y el de Peutz-Jeghers, que fueron detectados en una muestra con un fenotipo ultrahipermutado a nivel somático (*SEC23B* y *STK11IP*) (D. P. Smith et al., 2001; Yehia et al., 2015). Un total de 10 genes con SNV/indel germinal y LOH somática fueron así priorizados por estas dos estrategias, reparación del ADN y síndromes de cáncer hereditario, que unidos a los dos genes con SNV germinal y SNV somática hacen un total de 12 candidatos a tener en cuenta para la predisposición al CCR familiar (**Figura 21**). Adicionalmente se realizó un análisis de casos y controles para estos genes utilizando los datos de 1.006 casos de CCR de aparición temprana de la base de datos CanVar (Chubb, Broderick, Dobbins, & Houlston, 2016), así como la base de datos ExAC como controles (Lek et al., 2016), obteniendo un

enriquecimiento en casos para las variantes afectando a los genes *ADCY8*, *BLM*, *BRCA2*, *ERCC2*, *REV3L*, *RIF1*, *SEC23*, *SMARCA4* y *STK11IP*. También se realizó una caracterización mutacional somática a través de la aplicación MuSiCa desarrollada en el primer estudio de esta tesis (Díaz-Gay et al., 2018), con el objetivo de añadir más evidencia de cara a la priorización de candidatos a la predisposición al CCR familiar. Se evaluaron tanto la TMB como las contribuciones de las firmas mutacionales según las firmas de referencia v2 de COSMIC (Wellcome Trust Sanger Institute, 2019a). Cabe destacar que se encontraron un total de cinco tumores hipermutados, lo que está de acuerdo con el enriquecimiento encontrado previamente en funciones relacionadas con la reparación del ADN entre los candidatos seleccionados (Campbell et al., 2017).

Respecto a los 12 candidatos inicialmente seleccionados, los dos que presentaban SNV germinal y somática (*ADCY8* y *HSPG2*) fueron descartados para análisis posteriores al identificarse un rol potencialmente oncogénico después de un curado funcional más exhaustivo (mientras que el modelo de Knudson está basado en GSTs) (Hong et al., 2013; B. Sharma et al., 1998). Por otro lado, entre los genes con SNV/indel germinal e inactivación somática predicha por LOH, finalmente se destacaron seis genes, incluyendo los genes de predisposición ya conocidos para otras neoplasias *BLM*, *BRCA2* y *ERCC2*, así como los genes asociados a la reparación de ADN *RECQL*, *REV3L* y *RIF1*.

BLM y *RECQL* pertenecen ambos a la familia RecQ de helicasas, responsables de la apertura del ADN de doble cadena y con funciones en replicación, recombinación, transcripción y reparación del ADN (Croteau et al., 2014). Cabe destacar que mutaciones germinales bialélicas en *BLM* causan el síndrome de cáncer hereditario de Bloom (Ellis et al., 1995), mientras que en el caso de *RECQL*, las variantes encontradas pertenecen a un paciente en el que se encontró un fenotipo hipermutado (cerca de 100 mutaciones por megabase secuenciada) en el tumor. Ambos genes han sido además propuestos recientemente como genes de predisposición a cáncer de mama (Thompson et al., 2012; Cybulski et al., 2015), mientras que *BLM* ya había sido propuesto previamente para la predisposición al CCR (de Voer et al., 2015). *BRCA2* constituye uno de los genes hereditarios clásicos para cáncer de mama y ovario (Wooster et al., 1995), y en el caso de nuestra cohorte su doble alteración germinal-somática se ha encontrado en un paciente perteneciente a una familia con varios miembros también afectados por cáncer de mama. Así, ha sido seleccionado como el gen responsable del fenotipo en la familia, descartando por tanto a *PARP2*, que había sido detectado en el mismo paciente. Respecto a *ERCC2*, alteraciones germinales bialélicas causan *xeroderma pigmentosum*, un síndrome hereditario responsable de una susceptibilidad incrementada al cáncer de piel (Frederick et al., 1994). Este gen, perteneciente a la vía de reparación del ADN por escisión de nucleótidos, también se ha propuesto como candidato para predisposición a cáncer de mama y ovario (Rump et al., 2016). Por último, *REV3L* y *RIF1* se han asociado

a la reparación del ADN en dos vías diferenciadas, la de síntesis de ADN translesión y la reparación de roturas de doble cadena de ADN por unión de extremos no homólogos, respectivamente (Lange et al., 2011; Escribano-Díaz et al., 2013). Adicionalmente, se descartó el gen candidato *SMARCA4* (del que también se había predicho su inactivación en la misma familia que *REV3L*), después de evaluar la validación de LOH por secuenciación Sanger realizada en estudios previos (Esteban-Jurado et al., 2015, 2016). También cabe destacar que los genes *SEC23B* y *STK11IP*, detectados en una muestra con un tumor ultrahipermutado (más de 500 mutaciones por megabase), fueron finalmente descartados al esperarse un defecto en alguna vía de reparación del ADN en este caso debido al alto número de mutaciones encontrado.

Respecto al análisis de firmas mutacionales, se encontró una predominancia de la firma SBS1 asociada a la edad, lo que está de acuerdo con los análisis previos realizados con MuSiCa en la cohorte de cáncer de colon del TCGA, para las muestras sin IMS ni mutaciones en *POLE*. Sin embargo, esto estaría en desacuerdo con los altos valores de TMB encontrados en la cohorte, especialmente en los cinco casos hipermutados (Muzny et al., 2012). Cabe destacar también que no se encontró ninguna de las firmas asociadas a defectos en la reparación del ADN con una contribución significativa en el perfil mutacional de las muestras analizadas.

El análisis integrado germinal-tumoral desarrollado está de acuerdo con las recomendaciones recientes del Clinical Genome Resource, que propone el uso de la TMB y el análisis de firmas en la práctica clínica rutinaria. También considera la evaluación del segundo *hit* somático, aunque en este caso se recomienda un análisis caso por caso y bajo asesoramiento de un panel multidisciplinario de expertos en cada centro (Walsh et al., 2018). Cabe destacar que los potenciales candidatos a la predisposición germinal al CCR familiar identificados en este estudio podrían ser útiles en un futuro en la práctica clínica, permitiendo mejorar el diagnóstico en las familias afectadas. Sin embargo, la validación de las alteraciones genéticas encontradas mediante técnicas ortogonales, así como la replicación en cohortes independientes de CCR familiar y estudios funcionales serían necesarios para la confirmación de su asociación con el CCR hereditario, así como para proporcionar nuevo conocimiento acerca de los mecanismos moleculares implicados.

Conclusiones

1. *Mutational Signatures in Cancer* (MuSiCa) es una aplicación web de manejo sencillo y acceso libre desarrollada a través de la plataforma Shiny para realizar caracterización mutacional somática de muestras de cáncer.

2. MuSiCa se ha establecido como una de las aplicaciones web de referencia para el cálculo de la carga mutacional tumoral y la caracterización de las firmas mutacionales

según las firmas de referencia de COSMIC, siendo ampliamente utilizada desde su publicación.

3. La clasificación de muestras por *clustering* y análisis de componentes principales según las contribuciones de las distintas firmas mutacionales es una característica distintiva de MuSiCa, que no está disponible en ninguna de las aplicaciones competidoras que existen para realizar análisis de firmas mutacionales.

4. La caracterización molecular de muestras somáticas de CCR procedentes del proyecto TCGA se replicó de forma sencilla y precisa a través del análisis de firmas mutacionales de MuSiCa.

5. El análisis integrado de datos de SEC germinales y tumorales, teniendo en cuenta distintas clases de variantes genéticas y basado en la hipótesis clásica de los dos *hits* de Knudson y la caracterización mutacional somática, se ha demostrado útil para la identificación de nuevos GSTs candidatos a estar involucrados en la predisposición al CCR familiar.

6. Se identificaron seis genes como potenciales candidatos para la predisposición germinal al CCR familiar, incluyendo genes conocidos por su implicación en la predisposición a otras neoplasias, como es el caso de *BLM*, *BRCA2* y *ERCC2*, así como genes asociados a la reparación del ADN, *RECQL*, *REV3L* y *RIF1*.

7. El análisis del perfil mutacional somático puede ser útil en el descubrimiento del defecto germinal responsable. En nuestro estudio, esto se ejemplificó con un gen candidato ligado a la reparación del ADN, *RECQL*, que se encontró mutado en el ADN germinal de un paciente con un fenotipo hipermutado en el tumor, reforzando el rol potencial de este gen en el CCR hereditario.

APPENDICES

Appendix I. Publications associated with this thesis

Díaz-Gay, M., Vila-Casadesús, M., Franch-Expósito, S., Hernández-Illán, E., Lozano, J. J., & Castellví-Bel, S. (2018). Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics, 19*(1), 224. https://doi.org/10.1186/s12859-018-2234-y

Díaz-Gay, M., Franch-Expósito, S., Arnau-Collell, C., Park, S., Supek, F., Muñoz, J., … Castellví-Bel, S. (2019). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *Cancers, 11*(3), 362. https://doi.org/10.3390/cancers11030362

Appendix II. Publications in other collaboration projects

Original articles

Esteban-Jurado, C., Giménez-Zaragoza, D., Muñoz, J., Franch-Expósito, S., Álvarez-Barona, M., Ocaña, T., … Castellví-Bel, S. (2017). POLE and POLD1 screening in 155 patients with multiple polyps and early-onset colorectal cancer. *Oncotarget*, *8*(16). https://doi.org/10.18632/oncotarget.15810

Franch-Expósito, S., Esteban-Jurado, C., Garre, P., Quintanilla, I., Duran-Sanchon, S., **Díaz-Gay, M.**, … Castellví-Bel, S. (2018). Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis. *Journal of Genetics and Genomics*, *45*(1), 41–45. https://doi.org/10.1016/j.jgg.2017.12.001

Torabi, K., Erola, P., Álvarez-Mora, M. I., **Díaz-Gay, M.**, Ferrer, Q., Castells, A., … Camps, J. (2019). Quantitative analysis of somatically acquired and constitutive uniparental disomy in gastrointestinal cancers. *International Journal of Cancer*, *144*(3), 513–524. https://doi.org/10.1002/ijc.31936

Toma, C.*, **Díaz-Gay, M.***, Franch-Expósito, S., Arnau-Collell, C., Overs, B., Muñoz, J., … Castellví-Bel, S. (2019). Using linkage studies combined with whole-exome sequencing to identify novel candidate genes for familial colorectal cancer. *International Journal of Cancer*. https://doi.org/10.1002/ijc.32683

** shared first authors*

Toma, C.*, **Díaz-Gay, M.***, Soares de Lima, Y., Arnau-Collell, C., Franch-Expósito, S., Muñoz, J., … Castellví-Bel, S. (2019). Identification of a novel candidate gene for serrated polyposis syndrome by performing linkage analysis combined with whole-exome sequencing. *Clinical and Translational Gastroenterology*. *In press.*

Reviews

Grolleman, J. E.*, **Díaz-Gay, M.***, Franch-Expósito, S., Castellví-Bel, S., & de Voer, R. M. (2019). Somatic mutational signatures in polyposis and colorectal cancer. *Molecular Aspects of Medicine*, 69, 62–72. https://doi.org/10.1016/j.mam.2019.05.002

Appendix III. Conference communications associated with this thesis

International meetings

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2017). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *Human Genome Meeting 2017 (Human Genome Organisation)*. Barcelona (Spain): Oral communication.

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2017). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *EuCOLONGENE COST Management Committee and Working Groups Joint Meeting (COST Action BM1206)*. Porto (Portugal): Oral communication.

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2017). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *Translating Colorectal Cancer Research Workshop (COST Action BM1206)*. Porto (Portugal): Poster.

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2017). Integrated analysis of germline and tumor DNA for the identification of new candidate genes involved in familial colorectal cancer. *European Human Genetics Conference 2017 (European Society of Human Genetics)*. Copenhagen (Denmark): Poster.

Díaz-Gay, M., & Castellví-Bel, S. (2019). A new user-friendly web application to implement mutational signatures analysis. *31st European Congress of Pathology*. Nice (France): Invited oral communication.

National meetings

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2016). Identification of new genes involved in familial colorectal cancer: integrated analysis of germline and tumor DNA. *X Jornadas Científicas CIBERehd*. Barcelona (Spain): Poster. Best poster award.

Díaz-Gay, M., Franch-Expósito, S., Esteban-Jurado, C., Muñoz, J., Gratacós-Mulleras, A., Ocaña, T., … Castellví-Bel, S. (2016). Integrated analysis of germline and tumor DNA for the identification of new genes involved in familial colorectal cancer. *IV Bioinformatics and Genomics Symposium (Societat Catalana de Biologia / Bioinformatics Barcelona)*. Barcelona (Spain): Oral communication.

Díaz-Gay, M., Vila-Casadesús, M., Franch-Expósito, S., Lozano, J. J., & Castellví-Bel, S. (2017). Mutational Signatures in Cancer (MuSiC): a web application to implement mutational signatures framework in cancer samples. *V Bioinformatics and Genomics Symposium (Societat Catalana de Biologia / Bioinformatics Barcelona)*. Barcelona (Spain): Poster.

Díaz-Gay, M., Franch-Expósito, S., Park, S., Supek, F., Muñoz, J., Arnau-Collell, C., … Castellví-Bel, S. (2018). Integrated analysis of germline and tumor DNA for the identification of new candidate genes involved in familial colorectal cancer. *XII Jornadas Científicas CIBERehd*. Barcelona (Spain): Poster.

Díaz-Gay, M., Franch-Expósito, S., Park, S., Supek, F., Muñoz, J., Arnau-Collell, C., … Castellví-Bel, S. (2018). Integrated analysis of germline and tumor DNA for the identification of new candidate genes involved in familial colorectal cancer. *VI Bioinformatics and Genomics Symposium (Societat Catalana de Biologia / Bioinformatics Barcelona)*. Barcelona (Spain): Poster.

Díaz-Gay, M., Franch-Expósito, S., Park, S., Supek, F., Muñoz, J., Arnau-Collell, C., … Castellví-Bel, S. (2019). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *IDIBAPS PhD Day*. Barcelona (Spain): Poster.

BIBLIOGRAPHY

Adam, R., Spier, I., Zhao, B., Kloth, M., Marquez, J., Hinrichsen, I., … al., et. (2016). Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of colorectal adenomatous polyposis. *American Journal of Human Genetics*, *99*(2), 337–351. https://doi.org/10.1016/j.ajhg.2016.06.015

Ahadova, A., Gallon, R., Gebert, J., Ballhausen, A., Endris, V., Kirchner, M., … Kloor, M. (2018). Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. *International Journal of Cancer*, *143*(1), 139–150. https://doi.org/10.1002/ijc.31300

Al-Tassan, N., Chmiel, N. H., Maynard, J., Fleming, N., Livingston, A. L., Williams, G. T., … Cheadle, J. P. (2002). Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genetics*, *30*(2), 227–232. https://doi.org/10.1038/ng828

AlDubayan, S. H., Giannakis, M., Moore, N. D., Han, G. C., Reardon, B., Hamada, T., … Van Allen, E. M. (2018). Inherited DNA-repair defects in colorectal cancer. *American Journal of Human Genetics*, *102*(3), 401–414. https://doi.org/10.1016/j.ajhg.2018.01.018

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, *47*(12), 1402–1407. https://doi.org/10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Wu, Y., … Stratton, M. R. (2019). The repertoire of mutational signatures in human cancer. *BioRxiv*, 322859. https://doi.org/10.1101/322859

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V, … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, *3*(1), 246–259. https://doi.org/10.1016/j.celrep.2012.12.008

Algra, A. M., & Rothwell, P. M. (2012). Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *The Lancet Oncology*, *13*(5), 518–527. https://doi.org/10.1016/S1470-2045(12)70112-2

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M., … Lewis, C. (2018). Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, *391*(10125), 1023–1075. https://doi.org/10.1016/S0140-6736(17)33326-3

Aoude, L. G., Heitzer, E., Johansson, P., Gartside, M., Wadt, K., Pritchard, A. L., … Hayward, N. K. (2015). POLE mutations in families predisposed to cutaneous melanoma. *Familial Cancer*, *14*(4), 621–628. https://doi.org/10.1007/s10689-015-9826-8

Ardin, M., Cahais, V., Castells, X., Bouaoun, L., Byrnes, G., Herceg, Z., … Olivier, M. (2016). MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics*, *17*(1), 170. https://doi.org/10.1186/s12859-016-1011-z

Aronson, M., Gallinger, S., Cohen, Z., Cohen, S., Dvir, R., Elhasid, R., … Durno, C. (2016). Gastrointestinal findings in the largest series of patients with hereditary biallelic mismatch repair deficiency syndrome: report from the International Consortium. *American Journal of Gastroenterology*, *111*(2), 275–284. https://doi.org/10.1038/ajg.2015.392

Arora, S., Yan, H., Cho, I., Fan, H.-Y., Luo, B., Gai, X., … Enders, G. H. (2015). Genetic variants that predispose to DNA double-strand breaks in lymphocytes from a subset of patients with familial colorectal carcinomas. *Gastroenterology*, *149*(7), 1872-1883.e9. https://doi.org/10.1053/j.gastro.2015.08.052

Baez-Ortega, A., & Gori, K. (2019). Computational approaches for discovery of mutational signatures in cancer. *Briefings in Bioinformatics*, *20*(1), 77–88. https://doi.org/10.1093/bib/bbx082

Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., … La Vecchia, C. (2015). Alcohol consumption and site-specific cancer risk: a comprehensive dose–response meta-analysis. *British Journal of Cancer*, *112*(3), 580–593. https://doi.org/10.1038/bjc.2014.579

Bakry, D., Aronson, M., Durno, C., Rimawi, H., Farah, R., Alharbi, Q. K., … Tabori, U. (2014). Genetic and clinical determinants of constitutional mismatch repair deficiency syndrome: Report from the constitutional mismatch repair deficiency consortium. *European Journal of Cancer*, *50*(5), 987–996. https://doi.org/10.1016/j.ejca.2013.12.005

Bardou, M., Barkun, A., & Martel, M. (2010). Effect of statin therapy on colorectal cancer. *Gut*, *59*(11), 1572–1585. https://doi.org/10.1136/gut.2009.190900

Bashyam, M. D., Animireddy, S., Bala, P., Naz, A., & George, S. A. (2019). The Yin and Yang of cancer genes. *Gene*, *704*, 121–133. https://doi.org/10.1016/j.gene.2019.04.025

Bayati, M., Rabiee, H. R., Mehrbod, M., Vafaee, F., Ebrahimi, D., Forrest, A., & Alinejad-Rokny, H. (2019). CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes. *BioRxiv*, 424960. https://doi.org/10.1101/424960

Belhadj, S., Moutinho, C., Mur, P., Setien, F., Llinàs-Arias, P., Pérez-Salvia, M., … Valle, L. (2019). Germline variation in O6-methylguanine-DNA methyltransferase (MGMT) as cause of hereditary colorectal cancer. *Cancer Letters*, *447*, 86–92. https://doi.org/10.1016/j.canlet.2019.01.019

Bellido, F., Pineda, M., Aiza, G., Valdés-Mas, R., Navarro, M., Puente, D. A., … Valle, L. (2016). POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and/or polyposis: review of reported cases and recommendations for genetic testing and surveillance. *Genetics in Medicine*, *18*(4), 325–332. https://doi.org/10.1038/gim.2015.75

Bellido, F., Sowada, N., Mur, P., Lázaro, C., Pons, T., Valdés-Mas, R., … Valle, L. (2018). Association between germline mutations in BRF1, a subunit of the RNA polymerase III transcription complex, and hereditary colorectal cancer. *Gastroenterology*, *154*(1), 181-194.e20. https://doi.org/10.1053/j.gastro.2017.09.005

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, *52*(1), 155–173. https://doi.org/10.1016/j.csda.2006.11.006

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, *321*(6067), 209–213. https://doi.org/10.1038/321209a0

Biswas, S., Ellis, A. J., Guy, R., Savage, H., Madronal, K., & East, J. E. (2013). High prevalence of hyperplastic polyposis syndrome (serrated polyposis) in the NHS bowel cancer screening programme. *Gut*, *62*(3), 475. https://doi.org/10.1136/gutjnl-2012-303233

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., … van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, *538*(7624), 260–264. https://doi.org/10.1038/nature19768

Blokzijl, F., Janssen, R., van Boxtel, R., & Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine*, *10*(1), 33. https://doi.org/10.1186/s13073-018-0539-0

Bodmer, W. F., Bailey, C. J., Bodmer, J., Bussey, H. J. R., Ellis, A., Gorman, P., … Spurr, N. K. (1987). Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature*, *328*(6131), 614–616. https://doi.org/10.1038/328614a0

Borchers, H. W. (2019). *pracma: practical numerical math functions.* R package. Retrieved from https://cran.r-project.org/package=pracma

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B., & Maisonneuve, P. (2008). Smoking and colorectal cancer: a meta-analysis. *JAMA*, *300*(23), 2765–2778. https://doi.org/10.1001/jama.2008.839

Bouvard, V., Loomis, D., Guyton, K. Z., Grosse, Y., Ghissassi, F. El, Benbrahim-Tallaa, L., … Straif, K. (2015). Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology*, *16*(16), 1599–1600. https://doi.org/10.1016/S1470-2045(15)00444-1

Boyle, T., Keegel, T., Bull, F., Heyworth, J., & Fritschi, L. (2012). Physical activity and risks of proximal and distal colon cancers: a systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute*, *104*(20), 1548–1561. https://doi.org/10.1093/jnci/djs354

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Brea-Fernandez, A. J., Fernandez-Rozadilla, C., Alvarez-Barona, M., Azuara, D., Ginesta, M. M., Clofent, J., … Ruiz-Ponte, C. (2017). Candidate predisposing germline copy number variants in early onset colorectal cancer patients. *Clinical and Translational Oncology*, *19*(5), 625–632. https://doi.org/10.1007/s12094-016-1576-z

Brenner, H., & Chen, C. (2018). The colorectal cancer epidemic: challenges and opportunities for primary, secondary and tertiary prevention. *British Journal of Cancer*, *119*(7), 785–792. https://doi.org/10.1038/s41416-018-0264-x

Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *The Lancet*, *383*(9927), 1490–1502. https://doi.org/10.1016/S0140-6736(13)61649-9

Brockschmidt, A., Trost, D., Peterziel, H., Zimmermann, K., Ehrler, M., Grassmann, H., … Weber, R. G. (2012). KIAA1797/FOCAD encodes a novel focal adhesion protein with tumour suppressor function in gliomas. *Brain*, *135*(4), 1027–1041. https://doi.org/10.1093/brain/aws045

Broderick, P., Dobbins, S. E., Chubb, D., Kinnersley, B., Dunlop, M. G., Tomlinson, I., & Houlston, R. S. (2017). Validation of recently proposed colorectal cancer susceptibility gene variants in an analysis of families and patients - a systematic review. *Gastroenterology*, *152*(1), 75-77.e4. https://doi.org/10.1053/j.gastro.2016.09.041

Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., … Liskay, R. M. (1994). Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer. *Nature*, *368*(6468), 258–261. https://doi.org/10.1038/368258a0

Bruna, A., Rueda, O. M., Greenwood, W., Batra, A. S., Callari, M., Batra, R. N., … Caldas, C. (2016). A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell*, *167*(1), 260-274.e22. https://doi.org/10.1016/j.cell.2016.08.041

Buchanan, D. D., Clendenning, M., Zhuoer, L., Stewart, J. R., Joseland, S., Woodall, S., … Rosty, C. (2017). Lack of evidence for germline RNF43 mutations in patients with serrated polyposis syndrome from a large multinational study. *Gut*, *66*(6), 1170–1172. https://doi.org/10.1136/gutjnl-2016-312773

Buchanan, D. D., Sweet, K., Drini, M., Jenkins, M. A., Win, A. K., English, D. R., … Young, J. P.

(2010). Risk factors for colorectal cancer in patients with multiple serrated polyps: a cross-sectional case series from genetics clinics. *PLoS ONE*, *5*(7), e11636. https://doi.org/10.1371/journal.pone.0011636

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bussey, H. J. R. (1975). *Familial polyposis coli: family studies, histopathology, differential diagnosis, and results of treatment*. Baltimore: Johns Hopkins University Press.

Campbell, B. B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., … Shlien, A. (2017). Comprehensive analysis of hypermutation in human cancer. *Cell*, *171*(5), 1042-1056.e10. https://doi.org/10.1016/j.cell.2017.09.048

Carballal, S., Moreira, L., & Balaguer, F. (2013). Pólipos serrados y síndrome de poliposis serrada. *Cirugía Española*, *91*(3), 141–148. https://doi.org/10.1016/j.ciresp.2012.12.001

Carballal, S., Rodríguez-Alcalde, D., Moreira, L., Hernández, L., Rodríguez, L., Rodríguez-Moranta, F., … Balaguer, F. (2016). Colorectal cancer risk factors in patients with serrated polyposis syndrome: a large multicentre study. *Gut*, *65*(11), 1829–1837. https://doi.org/10.1136/gutjnl-2015-309647

Carethers, J. M., & Jung, B. H. (2015). Genetics and genetic biomarkers in sporadic colorectal cancer. *Gastroenterology*, *149*(5), 1177–1190. https://doi.org/10.1053/j.gastro.2015.06.047

Carlson, J., Li, J. Z., & Zöllner, S. (2018). Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. *BMC Genomics*, *19*(1), 845. https://doi.org/10.1186/s12864-018-5264-y

Castellsagué, E., Li, R., Aligue, R., González, S., Sanz, J., Martin, E., … Foulkes, W. D. (2019). Novel POLE pathogenic germline variant in a family with multiple primary tumors results in distinct mutational signatures. *Human Mutation*, *40*(1), 36–41. https://doi.org/10.1002/humu.23676

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., … Frampton, G. M. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, *9*(1), 34. https://doi.org/10.1186/s13073-017-0424-2

Chan, T. A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., & Peters, S. (2018). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology*, *30*(1), 44–56. https://doi.org/10.1093/annonc/mdy495

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, *18*(8), 495–506. https://doi.org/10.1038/nrm.2017.48

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2019). *shiny: web application framework for R.* R package. Retrieved from https://cran.r-project.org/package=shiny

Chauveau, C., Rowell, J., & Ferreiro, A. (2014). A rising titan: TTN review and mutation update. *Human Mutation*, *35*(9), 1046–1059. https://doi.org/10.1002/humu.22611

Chen, Z., Wen, W., Beeghly-Fadiel, A., Shu, X., Díez-Obrero, V., Long, J., … Guo, X. (2019). Identifying putative susceptibility genes and evaluating their associations with somatic mutations in human cancers. *American Journal of Human Genetics*, *105*(3), 477–492. https://doi.org/10.1016/j.ajhg.2019.07.006

Chester, N., Babbe, H., Pinkas, J., Manning, C., & Leder, P. (2006). Mutation of the murine Bloom's syndrome gene produces global genome destabilization. *Molecular and Cellular*

Biology, *26*(17), 6713–6726. https://doi.org/10.1128/MCB.00296-06

Cho, K. R., & Vogelstein, B. (1992). Genetic alterations in the adenoma–carcinoma sequence. *Cancer*, *70*(6 Suppl), 1727–1731. https://doi.org/10.1002/1097-0142(19920915)70:4+<1727::aid-cncr2820701613>3.0.co;2-p

Chubb, D., Broderick, P., Dobbins, S. E., Frampton, M., Kinnersley, B., Penegar, S., … Houlston, R. S. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature Communications*, *7*(1), 11883. https://doi.org/10.1038/ncomms11883

Chubb, D., Broderick, P., Dobbins, S. E., & Houlston, R. S. (2016). CanVar: A resource for sharing germline variation in cancer patients. *F1000Research*, *5*, 2813. https://doi.org/10.12688/f1000research.10058.1

Church, J. M., Hernegger, G. S., Moore, H. G., & Guillem, J. G. (2002). Attenuated familial adenomatous polyposis: an evolving and poorly understood entity. *Diseases of the Colon & Rectum*, *45*(1), 127–134. https://doi.org/10.1007/s10350-004-6127-y

Collins, A. R. (2007). Linkage disequilibrium and association mapping: an introduction. In A. R. Collins (Ed.), *Linkage disequilibrium and association mapping* (pp. 1–15). https://doi.org/10.1007/978-1-59745-389-9_1

Compe, E., & Egly, J.-M. (2012). TFIIH: when transcription met DNA repair. *Nature Reviews Molecular Cell Biology*, *13*(6), 343–354. https://doi.org/10.1038/nrm3350

Croteau, D. L., Popuri, V., Opresko, P. L., & Bohr, V. A. (2014). Human RecQ helicases in DNA repair, recombination, and replication. *Annual Review of Biochemistry*, *83*(1), 519–552. https://doi.org/10.1146/annurev-biochem-060713-035428

Cunningham, D., Atkin, W., Lenz, H. J., Lynch, H. T., Minsky, B., Nordlinger, B., & Starling, N. (2010). Colorectal cancer. *The Lancet*, *375*(9719), 1030–1047. https://doi.org/10.1016/S0140-6736(10)60353-4

Cybulski, C., Carrot-Zhang, J., Kluźniak, W., Rivera, B., Kashyap, A., Wokołorczyk, D., … Akbari, M. R. (2015). Germline RECQL mutations are associated with breast cancer susceptibility. *Nature Genetics*, *47*(6), 643–646. https://doi.org/10.1038/ng.3284

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., … Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, *23*(4), 517–525. https://doi.org/10.1038/nm.4292

de Voer, R. M., Geurts van Kessel, A., Weren, R. D. A., Ligtenberg, M. J. L., Smeets, D., Fu, L., … Kuiper, R. P. (2013). Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. *Gastroenterology*, *145*(3), 544–547. https://doi.org/10.1053/j.gastro.2013.06.001

de Voer, R. M., Hahn, M.-M., Mensenkamp, A. R., Hoischen, A., Gilissen, C., Henkes, A., … Kuiper, R. P. (2015). Deleterious germline BLM mutations and the risk for early-onset colorectal cancer. *Scientific Reports*, *5*(1), 14060. https://doi.org/10.1038/srep14060

de Voer, R. M., Hahn, M.-M., Weren, R. D. A., Mensenkamp, A. R., Gilissen, C., van Zelst-Stams, W. A., … Kuiper, R. P. (2016). Identification of novel candidate genes for early-onset colorectal cancer susceptibility. *PLOS Genetics*, *12*(2), e1005880. https://doi.org/10.1371/journal.pgen.1005880

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

DeRycke, M. S., Gunawardena, S. R., Middha, S., Asmann, Y. W., Schaid, D. J., McDonnell, S. K., … Goode, E. L. (2013). Identification of novel variants in colorectal cancer families by high-

throughput exome sequencing. *Cancer Epidemiology Biomarkers and Prevention*, *22*(7), 1239–1251. https://doi.org/10.1158/1055-9965.EPI-12-1226

Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J., Hillenmeyer, M. E., Davis, R. W., … Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, *169*(4), 1915–1925. https://doi.org/10.1534/genetics.104.036871

Devaraj, B., Lee, A., Cabrera, B. L., Miyai, K., Luo, L., Ramamoorthy, S., … Carethers, J. M. (2010). Relationship of EMAST and microsatellite instability among patients with rectal cancer. *Journal of Gastrointestinal Surgery*, *14*(10), 1521–1528. https://doi.org/10.1007/s11605-010-1340-6

Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Computational Biology*, *4*(7), e1000029. https://doi.org/10.1371/journal.pcbi.1000029

Díaz-Gay, M., Franch-Expósito, S., Arnau-Collell, C., Park, S., Supek, F., Muñoz, J., … Castellví-Bel, S. (2019). Integrated analysis of germline and tumor DNA identifies new candidate genes involved in familial colorectal cancer. *Cancers*, *11*(3), 362. https://doi.org/10.3390/cancers11030362

Díaz-Gay, M., Vila-Casadesús, M., Franch-Expósito, S., Hernández-Illán, E., Lozano, J. J., & Castellví-Bel, S. (2018). Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics*, *19*(1), 224. https://doi.org/10.1186/s12859-018-2234-y

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., & Tabernero, J. (2017). Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer*, *17*(4), 268–268. https://doi.org/10.1038/nrc.2017.24

Domingo, J. L., & Nadal, M. (2017). Carcinogenicity of consumption of red meat and processed meat: A review of scientific news since the IARC decision. *Food and Chemical Toxicology*, *105*, 256–261. https://doi.org/10.1016/j.fct.2017.04.028

Drost, J., van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., … Clevers, H. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science*, *358*(6360), 234–238. https://doi.org/10.1126/science.aao3130

Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., … Bass, A. J. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genetics*, *45*(5), 478–486. https://doi.org/10.1038/ng.2591

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Eggers, S., Smith, K. R., Bahlo, M., Looijenga, L. H., Drop, S. L., Juniarto, Z. A., … Sinclair, A. H. (2015). Whole exome sequencing combined with linkage analysis identifies a novel 3 bp deletion in NR5A1. *European Journal of Human Genetics*, *23*(4), 486–493. https://doi.org/10.1038/ejhg.2014.130

Ellis, N. A., Groden, J., Ye, T.-Z., Straughen, J., Lennon, D. J., Ciocci, S., … German, J. (1995). The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell*, *83*(4), 655–666. https://doi.org/10.1016/0092-8674(95)90105-1

Elsayed, F. a, Kets, C. M., Ruano, D., van den Akker, B., Mensenkamp, A. R., Schrumpf, M., … van Wezel, T. (2015). Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *European Journal of Human Genetics*, *23*(8), 1080–1084. https://doi.org/10.1038/ejhg.2014.242

Escribano-Díaz, C., Orthwein, A., Fradet-Turcotte, A., Xing, M., Young, J. T. F., Tkáč, J., …

Durocher, D. (2013). A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Molecular Cell*, *49*(5), 872–883. https://doi.org/10.1016/j.molcel.2013.01.001

Esteban-Jurado, C., Franch-Expósito, S., Muñoz, J., Ocaña, T., Carballal, S., López-Cerón, M., … Castellví-Bel, S. (2016). The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *European Journal of Human Genetics*, *24*(10), 1501–1505. https://doi.org/10.1038/ejhg.2016.44

Esteban-Jurado, C., Vila-Casadesús, M., Garre, P., Lozano, J. J., Pristoupilova, A., Beltran, S., … Castellví-Bel, S. (2015). Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genetics in Medicine*, *17*(2), 131–142. https://doi.org/10.1038/gim.2014.89

Evans, D. R., Venkitachalam, S., Revoredo, L., Dohey, A. T., Clarke, E., Pennell, J. J., … Guda, K. (2018). Evidence for GALNT12 as a moderate penetrance gene for colorectal cancer. *Human Mutation*, *39*(8), 1092–1101. https://doi.org/10.1002/humu.23549

Fantini, D., Glaser, A. P., Rimar, K. J., Wang, Y., Schipma, M., Varghese, N., … Meeks, J. J. (2018). A carcinogen-induced mouse model recapitulates the molecular alterations of human muscle invasive bladder cancer. *Oncogene*, *37*(14), 1911–1925. https://doi.org/10.1038/s41388-017-0099-6

Fearon, E. R. (2011). Molecular genetics of colorectal cancer. *Annual Review of Pathology: Mechanisms of Disease*, *6*(1), 479–507. https://doi.org/10.1146/annurev-pathol-011110-130235

Feinberg, A. P., Gehrke, C. W., Kuo, K. C., & Ehrlich, M. (1988). Reduced genomic 5-methylcytosine content in human colonic neoplasia. *Cancer Research*, *48*(5), 1159–1161. Retrieved from http://cancerres.aacrjournals.org/content/48/5/1159.abstract

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., … Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, *144*(8), 1941–1953. https://doi.org/10.1002/ijc.31937

Fievet, A., Mouret-Fourme, E., Colas, C., de Pauw, A., Stoppa-Lyonnet, D., & Buecher, B. (2019). Prevalence of pathogenic variants of FAN1 in more than 5000 patients assessed for genetic predisposition to colorectal, breast, ovarian, or other cancers. *Gastroenterology*, *156*(6), 1919–1920. https://doi.org/10.1053/j.gastro.2019.01.003

Fischer, A., Illingworth, C. J. R., Campbell, P. J., & Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, *14*(4), R39. https://doi.org/10.1186/gb-2013-14-4-r39

Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., Jenkins, N. A., Garber, J., … Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, *75*(5), 1027–1038. https://doi.org/10.1016/0092-8674(93)90546-3

Franch-Expósito, S., Esteban-Jurado, C., Garre, P., Quintanilla, I., Duran-Sanchon, S., Díaz-Gay, M., … Castellví-Bel, S. (2018). Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis. *Journal of Genetics and Genomics*, *45*(1), 41–45. https://doi.org/10.1016/j.jgg.2017.12.001

Frank, C., Sundquist, J., Yu, H., Hemminki, A., & Hemminki, K. (2017). Concordant and discordant familial cancer: Familial risks, proportions and population impact. *International Journal of Cancer*, *140*(7), 1510–1516. https://doi.org/10.1002/ijc.30583

Frederick, G. D., Amirkhan, R. H., Schultz, R. A., & Friedberg, E. C. (1994). Structural and mutational analysis of the xeroderma pigmentosum group D (XPD) gene. *Human*

Molecular Genetics, *3*(10), 1783–1788. https://doi.org/10.1093/hmg/3.10.1783

Funnell, T., Zhang, A. W., Grewal, D., McKinney, S., Bashashati, A., Wang, Y. K., & Shah, S. P. (2019). Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLOS Computational Biology*, *15*(2), e1006799. https://doi.org/10.1371/journal.pcbi.1006799

Gala, M. K., Mizukami, Y., Le, L. P., Moriichi, K., Austin, T., Yamamoto, M., … Chung, D. C. (2014). Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. *Gastroenterology*, *146*(2), 520-529.e6. https://doi.org/10.1053/j.gastro.2013.10.045

Garre, P., Martín, L., Sanz, J., Romero, A., Tosar, A., Bando, I., … Caldés, T. (2015). BRCA2 gene: a candidate for clinical testing in familial colorectal cancer type X. *Clinical Genetics*, *87*(6), 582–587. https://doi.org/10.1111/cge.12427

Gehring, J. S., Fischer, B., Lawrence, M., & Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, *31*(22), 3673–3575. https://doi.org/10.1093/bioinformatics/btv408

Giardiello, F. M., Welsh, S. B., Hamilton, S. R., Offerhaus, G. J. A., Gittelsohn, A. M., Booker, S. V., … Luk, G. D. (1987). Increased risk of cancer in the Peutz–Jeghers syndrome. *New England Journal of Medicine*, *316*(24), 1511–1514. https://doi.org/10.1056/NEJM198706113162404

Goel, A., & Boland, C. R. (2012). Epigenetics of colorectal cancer. *Gastroenterology*, *143*(6), 1442-1460.e1. https://doi.org/10.1053/j.gastro.2012.09.032

Goh, G., Schmid, R., Guiver, K., Arpornwirat, W., Chitapanarux, I., Ganju, V., … Swanton, C. (2016). Clonal evolutionary analysis during HER2 blockade in HER2-positive inflammatory breast cancer: a phase II open-label clinical trial of afatinib +/- vinorelbine. *PLOS Medicine*, *13*(12), e1002136. https://doi.org/10.1371/journal.pmed.1002136

Goldberg, Y., Halpern, N., Hubert, A., Adler, S. N., Cohen, S., Plesser-Duvdevani, M., … Meiner, V. (2015). Mutated MCM9 is associated with predisposition to hereditary mixed polyposis and colorectal cancer in addition to primary ovarian failure. *Cancer Genetics*, *208*(12), 621–624. https://doi.org/10.1016/j.cancergen.2015.10.001

Goncearenco, A., Rager, S. L., Li, M., Sang, Q.-X., Rogozin, I. B., & Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Research*, *45*(W1), W514–W522. https://doi.org/10.1093/nar/gkx367

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gori, K., & Baez-Ortega, A. (2018). sigfit: flexible Bayesian inference of mutational signatures. *BioRxiv*, 372896. https://doi.org/10.1101/372896

Grilley, M., Holmes, J., Yashar, B., & Modrich, P. (1990). Mechanisms of DNA-mismatch correction. *Mutation Research/DNA Repair*, *236*(2–3), 253–267. https://doi.org/10.1016/0921-8777(90)90009-T

Grolleman, J. E., de Voer, R. M., Elsayed, F. A., Nielsen, M., Weren, R. D. A., Palles, C., … Kuiper, R. P. (2019). Mutational signature analysis reveals NTHL1 deficiency to cause a multi-tumor phenotype. *Cancer Cell*, *35*(2), 256-266.e5. https://doi.org/10.1016/j.ccell.2018.12.011

Grolleman, J. E., Díaz-Gay, M., Franch-Expósito, S., Castellví-Bel, S., & de Voer, R. M. (2019). Somatic mutational signatures in polyposis and colorectal cancer. *Molecular Aspects of Medicine*, *69*, 62–72. https://doi.org/10.1016/j.mam.2019.05.002

Gruber, S. B., Ellis, N. A., Scott, K. K., Almog, R., Kolachana, P., Bonner, J. D., … Offit, K. (2002). BLM heterozygosity and the risk of colorectal cancer. *Science*, *297*(5589), 2013. https://doi.org/10.1126/science.1074399

Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., … Markowitz, S. D. (2009). Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proceedings of the National Academy of Sciences*, *106*(31), 12921–12925. https://doi.org/10.1073/pnas.0901454106

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., … Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350–1356. https://doi.org/10.1038/nm.3967

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., & Shyr, Y. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed Research International*, *2013*, 915636. https://doi.org/10.1155/2013/915636

Gupta, S., Provenzale, D., Llor, X., Halverson, A. L., Grady, W., Chung, D. C., … Ogba, N. (2019). NCCN guidelines insights: genetic/familial high-risk assessment: colorectal, version 2.2019. *Journal of the National Comprehensive Cancer Network*, *17*(9), 1032–1041. https://doi.org/10.6004/jnccn.2019.0044

Gylfe, A. E., Katainen, R., Kondelin, J., Tanskanen, T., Cajuso, T., Hänninen, U., … Aaltonen, L. A. (2013). Eleven candidate susceptibility genes for common familial colorectal cancer. *PLoS Genetics*, *9*(10), e1003876. https://doi.org/10.1371/journal.pgen.1003876

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hanane, O., Gianluca, S., & Vittorio, P. (2019). Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *BioRxiv*, 483982. https://doi.org/10.1101/483982

Hansen, M. F., Johansen, J., Bjørnevoll, I., Sylvander, A. E., Steinsbekk, K. S., Sætrom, P., … Sjursen, W. (2015). A novel POLE mutation associated with cancers of colon, pancreas, ovaries and small intestine. *Familial Cancer*, *14*(3), 437–448. https://doi.org/10.1007/s10689-015-9803-2

Hansen, M. F., Johansen, J., Sylvander, A. E., Bjørnevoll, I., Talseth-Palmer, B. A., Lavik, L. A. S., … Sjursen, W. (2017). Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. *Clinical Genetics*, *92*(4), 405–414. https://doi.org/10.1111/cge.12994

Hao, J.-J., Lin, D.-C., Dinh, H. Q., Mayakonda, A., Jiang, Y.-Y., Chang, C., … Koeffler, H. P. (2016). Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics*, *48*(12), 1500–1507. https://doi.org/10.1038/ng.3683

Haradhvala, N. J., Kim, J., Maruvka, Y. E., Polak, P., Rosebrock, D., Livitz, D., … Getz, G. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications*, *9*(1), 1746. https://doi.org/10.1038/s41467-018-04002-4

Hayward, N. K., Wilmott, J. S., Waddell, N., Johansson, P. A., Field, M. A., Nones, K., … Mann, G. J. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, *545*(7653), 175–180. https://doi.org/10.1038/nature22071

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in

human cancers. *Nature Reviews Genetics*, *15*(9), 585–598. https://doi.org/10.1038/nrg3729

Henrichsen, C. N., Chaignat, E., & Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics*, *18*(R1), 1–8. https://doi.org/10.1093/hmg/ddp011

Hermsen, M., Postma, C., Baak, J., Weiss, M., Rapallo, A., Sciutto, A., … Meijer, G. (2002). Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology*, *123*(4), 1109–1119. https://doi.org/10.1053/gast.2002.36051

Hino, O., & Kobayashi, T. (2017). Mourning Dr. Alfred G. Knudson: the two-hit hypothesis, tumor suppressor genes, and the tuberous sclerosis complex. *Cancer Science*, *108*(1), 5–11. https://doi.org/10.1111/cas.13116

Hong, S.-H., Goh, S.-H., Lee, S.-J., Hwang, J.-A., Lee, J., Choi, I.-J., … Lee, Y.-S. (2013). Upregulation of adenylate cyclase 3 (ADCY3) increases the tumorigenic potential of cells by activating the CREB pathway. *Oncotarget*, *4*(10), 1791–1803. https://doi.org/10.18632/oncotarget.1324

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), 3156. https://doi.org/10.1186/gb-2013-14-10-r115

Howe, J. R., Bair, J. L., Sayed, M. G., Anderson, M. E., Mitros, F. A., Petersen, G. M., … Vogelstein, B. (2001). Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nature Genetics*, *28*(2), 184–187. https://doi.org/10.1038/88919

Howe, J. R., Roth, S., Ringold, J. C., Summers, R. W., Järvinen, H. J., Sistonen, P., … Aaltonen, L. A. (1998). Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science*, *280*(5366), 1086–1088. https://doi.org/10.1126/science.280.5366.1086

Huang, P.-J., Chiu, L.-Y., Lee, C.-C., Yeh, Y.-M., Huang, K.-Y., Chiu, C.-H., & Tang, P. (2018). mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Research*, *46*(D1), D964–D970. https://doi.org/10.1093/nar/gkx1133

Huang, X., Wojtowicz, D., & Przytycka, T. M. (2018). Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, *34*(2), 330–337. https://doi.org/10.1093/bioinformatics/btx604

Huebschmann, D., Gu, Z., & Schlesner, M. (2019). *YAPSA: Yet Another Package for Signature Analysis.* R package. https://doi.org/10.18129/B9.bioc.YAPSA

IJspeert, J. E. G., Rana, S. A. Q., Atkinson, N. S. S., van Herwaarden, Y. J., Bastiaansen, B. A. J., van Leerdam, M. E., … Dekker, E. (2017). Clinical risk factors of colorectal cancer in patients with serrated polyposis syndrome: a multicentre cohort analysis. *Gut*, *66*(2), 278–284. https://doi.org/10.1136/gutjnl-2015-310630

IJspeert, J. E. G., Vermeulen, L., Meijer, G. a., & Dekker, E. (2015). Serrated neoplasia—role in colorectal carcinogenesis and clinical implications. *Nature Reviews Gastroenterology & Hepatology*, *12*(7), 401–409. https://doi.org/10.1038/nrgastro.2015.73

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., & Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, *363*(6429), 558–561. https://doi.org/10.1038/363558a0

Jaeger, E., Leedham, S., Lewis, A., Segditsas, S., Becker, M., Cuadrado, P. R., … Tomlinson, I. (2012). Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. *Nature Genetics*, *44*(6), 699–703. https://doi.org/10.1038/ng.2263

Jansen, A. M., van Wezel, T., van den Akker, B. E., Ventayol Garcia, M., Ruano, D., Tops, C. M., … Morreau, H. (2016). Combined mismatch repair and POLE/POLD1 defects explain unresolved suspected Lynch syndrome cancers. *European Journal of Human Genetics*,

24(7), 1089–1092. https://doi.org/10.1038/ejhg.2015.252

Jasperson, K. W., Kanth, P., Kirchhoff, A. C., Huismann, D., Gammon, A., Kohlmann, W., … Samadder, N. J. (2013). Serrated polyposis: colonic phenotype, extracolonic features, and familial risk in a large cohort. *Diseases of the Colon and Rectum*, *56*(11), 1211–1216. https://doi.org/10.1097/DCR.0b013e3182a11cca

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology*, *138*(6), 2044–2058. https://doi.org/10.1053/j.gastro.2010.01.054

Jelinic, P., Mueller, J. J., Olvera, N., Dao, F., Scott, S. N., Shah, R., … Levine, D. A. (2014). Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. *Nature Genetics*, *46*(5), 424–426. https://doi.org/10.1038/ng.2922

Jess, T., Rungoe, C., & Peyrin–Biroulet, L. (2012). Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies. *Clinical Gastroenterology and Hepatology*, *10*(6), 639–645. https://doi.org/10.1016/j.cgh.2012.01.010

Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B. J., … Hsu, L. (2014). Estimating the heritability of colorectal cancer. *Human Molecular Genetics*, *23*(14), 3898–3905. https://doi.org/10.1093/hmg/ddu087

Jozwiak, J., Jozwiak, S., Wlodarski, P., Consortium, E. C. 16, Slegtenhorst, M. van, Hoogt, R. de, … al., et. (2008). Possible mechanisms of disease development in tuberous sclerosis. *The Lancet Oncology*, *9*(1), 73–79. https://doi.org/10.1016/S1470-2045(07)70411-4

Kadalayil, L., Rafiq, S., Rose-Zerilli, M. J. J., Pengelly, R. J., Parker, H., Oscier, D., … Collins, A. (2015). Exome sequence read depth methods for identifying copy number changes. *Briefings in Bioinformatics*, *16*(3), 380–392. https://doi.org/10.1093/bib/bbu027

Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., … Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

Kanth, P., Grimmett, J., Champine, M., Burt, R., & Samadder, J. N. (2017). Hereditary colorectal polyposis and cancer syndromes: a primer on diagnosis and management. *American Journal of Gastroenterology*, *112*(10), 1509–1525. https://doi.org/10.1038/ajg.2017.212

Kanu, N., Cerone, M. A., Goh, G., Zalmas, L.-P., Bartkova, J., Dietzen, M., … Swanton, C. (2016). DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biology*, *17*(1), 185. https://doi.org/10.1186/s13059-016-1042-9

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, 531210. https://doi.org/10.1101/531210

Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., … Brown, J. R. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, *6*(1), 8866. https://doi.org/10.1038/ncomms9866

Khare, S., & Verma, M. (2012). Epigenetics of colon cancer. In R. G. Dumitrescu & M. Verma (Eds.), *Cancer epigenetics. Methods in molecular biology (methods and protocols)* (pp. 177–185). https://doi.org/10.1007/978-1-61779-612-8_10

Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Tiao, G., … Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, *48*(6), 600–606. https://doi.org/10.1038/ng.3557

Kinnersley, B., Chubb, D., Dobbins, S. E., Frampton, M., Buch, S., Timofeeva, M. N., … Houlston, R. S. (2016). Correspondence: SEMA4A variation and risk of colorectal cancer. *Nature Communications*, *7*, 10611. https://doi.org/10.1038/ncomms10611

Kinzler, K. W., & Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell*, *87*(2), 159–170. https://doi.org/10.1016/S0092-8674(00)81333-1

Kitao, S., Shimamoto, A., Goto, M., Miller, R. W., Smithson, W. A., Lindor, N. M., & Furuichi, Y. (1999). Mutations in RECQL4 cause a subset of cases of Rothmund-Thomson syndrome. *Nature Genetics*, *22*(1), 82–84. https://doi.org/10.1038/8788

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, *68*(4), 820–823. https://doi.org/10.1073/pnas.68.4.820

Krüger, S., & Piro, R. M. (2019). decompTumor2Sig: identification of mutational signatures active in individual tumors. *BMC Bioinformatics*, *20*(S4), 152. https://doi.org/10.1186/s12859-019-2688-6

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., … Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*(1), 15065. https://doi.org/10.1038/nrdp.2015.65

Lange, S. S., Takata, K., & Wood, R. D. (2011). DNA polymerases and cancer. *Nature Reviews Cancer*, *11*(2), 96–110. https://doi.org/10.1038/nrc2998

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell*, *177*(1), 70–84. https://doi.org/10.1016/j.cell.2019.02.032

Lauby-Secretan, B., Scoccianti, C., Loomis, D., Grosse, Y., Bianchini, F., & Straif, K. (2016). Body fatness and cancer — viewpoint of the IARC working group. *New England Journal of Medicine*, *375*(8), 794–798. https://doi.org/10.1056/NEJMsr1606602

Lawson, C. L., & Hanson, R. J. (1974). Solving least squares problems. In *Prentice-Hall series in automatic computation*. Englewood Cliffs: Prentice-Hall.

Leach, F. S., Nicolaides, N. C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., … Nyström-Lahti, M. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell*, *75*(6), 1215–1225. https://doi.org/10.1016/0092-8674(93)90330-s

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, J., Lee, A. J., Lee, J.-K., Park, J., Kwon, Y., Park, S., … Hong, D. (2018). Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Research*, *46*(W1), W102–W108. https://doi.org/10.1093/nar/gky406

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Leppert, M., Burt, R., Hughes, J. P., Samowitz, W., Nakamura, Y., Woodward, S., … White, R. (1990). Genetic analysis of an inherited predisposition to colon cancer in a family with a variable number of adenomatous polyps. *New England Journal of Medicine*, *322*(13), 904–908. https://doi.org/10.1056/NEJM199003293221306

Leppert, M., Dobbs, M., Scambler, P., O'Connell, P., Nakamura, Y., Stauffer, D., … Et, A. (1987). The gene for familial polyposis coli maps to the long arm of chromosome 5. *Science*, *238*(4832), 1411–1413. https://doi.org/10.1126/science.3479843

Li, J., Woods, S. L., Healey, S., Beesley, J., Chen, X., Lee, J. S., … Chenevix-Trench, G. (2016). Point mutations in exon 1B of APC reveal gastric adenocarcinoma and proximal polyposis of the stomach as a familial adenomatous polyposis variant. *American Journal of Human*

Genetics, *98*(5), 830–842. https://doi.org/10.1016/j.ajhg.2016.03.001

Liaw, D., Marsh, D. J., Li, J., Dahia, P. L. M., Wang, S. I., Zheng, Z., … Parsons, R. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nature Genetics*, *16*(1), 64–67. https://doi.org/10.1038/ng0597-64

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., … Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, *343*(2), 78–85. https://doi.org/10.1056/NEJM200007133430201

Ligtenberg, M. J. L., Kuiper, R. P., Chan, T. L., Goossens, M., Hebeda, K. M., Voorendt, M., … Hoogerbrugge, N. (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1. *Nature Genetics*, *41*(1), 112–117. https://doi.org/10.1038/ng.283

Limsui, D., Vierkant, R. A., Tillmans, L. S., Wang, A. H., Weisenberger, D. J., Laird, P. W., … Limburg, P. J. (2012). Postmenopausal hormone therapy and colorectal cancer risk by molecularly defined subtypes among older women. *Gut*, *61*(9), 1299–1305. https://doi.org/10.1136/gutjnl-2011-300719

Lindblom, A., Tannergård, P., Werelius, B., & Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non–polyposis colon cancer. *Nature Genetics*, *5*(3), 279–282. https://doi.org/10.1038/ng1193-279

Lindor, N. M., Rabe, K., Petersen, G. M., Haile, R., Casey, G., Baron, J., … Seminara, D. (2005). Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X. *JAMA*, *293*(16), 1979–1985. https://doi.org/10.1001/jama.293.16.1979

Loman, N. J., Misra, R. V, Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, *30*(5), 434–439. https://doi.org/10.1038/nbt.2198

Lynch, A. G. (2016). Decomposition of mutational context signatures using quadratic programming methods. *F1000Research*, *5*(1253), 1253. https://doi.org/10.12688/f1000research.8918.1

Lynch, H. T., & Krush, A. J. (1971). Cancer family "G" revisited: 1895-1970. *Cancer*, *27*(6), 1505–1511. https://doi.org/10.1002/1097-0142(197106)27:6<1505::AID-CNCR2820270635>3.0.CO;2-L

Lynch, H. T., Smyrk, T., Lanspa, S. J., Marcus, J. N., Kriegler, M., Lynch, J. F., & Appelman, H. D. (1988). Flat adenomas in a colon cancer-prone kindred. *JNCI Journal of the National Cancer Institute*, *80*(4), 278–282. https://doi.org/10.1093/jnci/80.4.278

Lynch, H. T., Smyrk, T., McGinn, T., Lanspa, S., Cavalieri, J., Lynch, J., … Luce, M. C. (1995). Attenuated familial adenomatous polyposis (AFAP) a phenotypically and genotypically distinctive variant of FAP. *Cancer*, *76*(12), 2427–2433. https://doi.org/10.1002/1097-0142(19951215)76:12<2427::AID-CNCR2820761205>3.0.CO;2-B

Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D., & Hitchins, M. P. (2015). Milestones of Lynch syndrome: 1895-2015. *Nature Reviews. Cancer*, *15*(3), 181–194. https://doi.org/10.1038/nrc3878

Ma, H., Brosens, L. A. A., Offerhaus, G. J. A., Giardiello, F. M., de Leng, W. W. J., & Montgomery, E. A. (2018). Pathology and genetics of hereditary colorectal cancer. *Pathology*, *50*(1), 49–59. https://doi.org/10.1016/j.pathol.2017.09.004

Ma, J., Setton, J., Lee, N. Y., Riaz, N., & Powell, S. N. (2018). The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nature Communications*, *9*(1), 3292. https://doi.org/10.1038/s41467-018-05228-y

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, *42*(D1), 986–992. https://doi.org/10.1093/nar/gkt958

Macintyre, G., Goranova, T. E., De Silva, D., Ennis, D., Piskorz, A. M., Eldridge, M., … Brenton, J. D. (2018). Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, *50*(9), 1262–1270. https://doi.org/10.1038/s41588-018-0179-8

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marteijn, J. A., Lans, H., Vermeulen, W., & Hoeijmakers, J. H. J. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, *15*(7), 465–481. https://doi.org/10.1038/nrm3822

Martín-Morales, L., Feldman, M., Vershinin, Z., Garre, P., Caldés, T., & Levy, D. (2017). SETD6 dominant negative mutation in familial colorectal cancer type X. *Human Molecular Genetics*, (September). https://doi.org/10.1093/hmg/ddx336

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

McDaniel, L. D., Chester, N., Watson, M., Borowsky, A. D., Leder, P., & Schultz, R. A. (2003). Chromosome instability and tumor predisposition inversely correlate with BLM protein levels. *DNA Repair*, *2*(12), 1387–1404. https://doi.org/10.1016/j.dnarep.2003.08.006

Mehenni, H., Lin-Marq, N., Buchet-Poyau, K., Reymond, A., Collart, M. A., Picard, D., & Antonarakis, S. E. (2005). LKB1 interacts with and phosphorylates PTEN: a functional link between two proteins involved in cancer predisposing syndromes. *Human Molecular Genetics*, *14*(15), 2209–2219. https://doi.org/10.1093/hmg/ddi225

Mendel, G. J. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen Des Naturforschenden Vereines in Brünn*, *4*, 3–47.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., … Mori, T. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genetics*, *17*(3), 271–272. https://doi.org/10.1038/ng1197-271

Moore, S. C., Lee, I.-M., Weiderpass, E., Campbell, P. T., Sampson, J. N., Kitahara, C. M., … Patel, A. V. (2016). Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. *JAMA Internal Medicine*, *176*(6), 816–825. https://doi.org/10.1001/jamainternmed.2016.1548

Moreira, L., Balaguer, F., Lindor, N., de la Chapelle, A., Hampel, H., Aaltonen, L. A., … EPICOLON Consortium. (2012). Identification of Lynch syndrome among patients with colorectal cancer. *JAMA*, *308*(15), 1555-1565. https://doi.org/10.1001/jama.2012.13088

Moreira, L., Pellisé, M., Carballal, S., Bessa, X., Ocaña, T., Serradesanferm, A., … Balaguer, F. (2013). High prevalence of serrated polyposis syndrome in FIT-based colorectal cancer screening programmes. *Gut*, *62*(3), 476–477. https://doi.org/10.1136/gutjnl-2012-303496

Mur, P., De Voer, R. M., Olivera-Salguero, R., Rodríguez-Perales, S., Pons, T., Setién, F., … Valle,

L. (2018). Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are infrequent in familial colorectal cancer and polyposis. *Molecular Cancer*, *17*(1), 1–6. https://doi.org/10.1186/s12943-018-0762-8

Murphy, N., Moreno, V., Hughes, D. J., Vodicka, L., Vodicka, P., Aglago, E. K., … Jenab, M. (2019). Lifestyle and dietary environmental factors in colorectal cancer susceptibility. *Molecular Aspects of Medicine*, *69*, 2–9. https://doi.org/10.1016/j.mam.2019.06.005

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., … Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

Nagahashi, M., Wakai, T., Shimada, Y., Ichikawa, H., Kameyama, H., Kobayashi, T., … Lyle, S. (2016). Genomic landscape of colorectal cancer in Japan: clinical implications of comprehensive genomic sequencing for precision medicine. *Genome Medicine*, *8*(1), 136. https://doi.org/10.1186/s13073-016-0387-8

Nagtegaal, I. D., Odze, R. D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., … Cree, I. A. (2019). The 2019 WHO classification of tumours of the digestive system. *Histopathology*. https://doi.org/10.1111/his.13975

Ngeow, J., Yu, W., Yehia, L., Niazi, F., Chen, J., Tang, X., … Eng, C. (2015). Exome sequencing reveals germline SMAD9 mutation that reduces phosphatase and tensin homolog expression and is associated with hamartomatous polyposis and gastrointestinal ganglioneuromas. *Gastroenterology*, *149*(4), 886–889. https://doi.org/10.1053/j.gastro.2015.06.027

Nicolaides, N. C., Papadopoulos, N., Liu, B., Wei, Y. F., Carter, K. C., Ruben, S. M., … Fraser, C. M. (1994). Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature*, *371*(6492), 75–80. https://doi.org/10.1038/371075a0

Nieminen, T. T., O'Donohue, M.-F., Wu, Y., Lohi, H., Scherer, S. W., Paterson, A. D., … Peltomäki, P. (2014). Germline mutation of RPS20, dncoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology*, *147*(3), 595-598.e5. https://doi.org/10.1053/j.gastro.2014.06.009

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–993. https://doi.org/10.1016/j.cell.2012.04.024

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Nordling, C. O. (1953). A new theory on the cancer-inducing mechanism. *British Journal of Cancer*, *7*(1), 68–72. https://doi.org/10.1038/bjc.1953.8

Norton, N., Li, D., Rampersaud, E., Morales, A., Martin, E. R., Zuchner, S., … Hershberger, R. E. (2013). Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy. *Circulation: Cardiovascular Genetics*, *6*(2), 144–153. https://doi.org/10.1161/circgenetics.111.000062

Okugawa, Y., Grady, W. M., & Goel, A. (2015). Epigenetic alterations in colorectal cancer: emerging biomarkers. *Gastroenterology*, *149*(5), 1204-1225.e12. https://doi.org/10.1053/j.gastro.2015.07.011

Olkinuora, A., Nieminen, T. T., Mårtensson, E., Rohlin, A., Ristimäki, A., Koskenvuo, L., … Peltomäki, P. (2019). Biallelic germline nonsense variant of MLH3 underlies polyposis predisposition. *Genetics in Medicine*, *21*(8), 1868–1873. https://doi.org/10.1038/s41436-018-0405-x

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, *16*(5), 275–284. https://doi.org/10.1038/nrg3908

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136–144. https://doi.org/10.1038/ng.2503

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., … Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, *11*(5), R52. https://doi.org/10.1186/gb-2010-11-5-r52

Papadopoulos, N., Nicolaides, N. C., Wei, Y. F., Ruben, S. M., Carter, K. C., Rosen, C. A., … Adams, M. D. (1994). Mutation of a mutL homolog in hereditary colon cancer. *Science*, *263*(5153), 1625–1629. https://doi.org/10.1126/science.8128251

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, *12*(4), 252–264. https://doi.org/10.1038/nrc3239

Park, S., Supek, F., & Lehner, B. (2018). Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. *Nature Communications*, *9*(1), 2601. https://doi.org/10.1038/s41467-018-04900-7

Parvathaneni, S., Stortchevoi, A., Sommers, J. A., Brosh, R. M., & Sharma, S. (2013). Human RECQ1 interacts with Ku70/80 and modulates DNA end-joining of double-strand breaks. *PloS One*, *8*(5), e62481. https://doi.org/10.1371/journal.pone.0062481

Peltomaki, P., Aaltonen, L., Sistonen, P., Pylkkanen, L., Mecklin, J., Jarvinen, H., … Et, A. (1993). Genetic mapping of a locus predisposing to human colorectal cancer. *Science*, *260*(5109), 810–812. https://doi.org/10.1126/science.8484120

Peters, U., Bien, S., & Zubair, N. (2015). Genetic architecture of colorectal cancer. *Gut*, *64*(10), 1623–1636. https://doi.org/10.1136/gutjnl-2013-306705

Pilarski, R. T., Brothman, A. R., Benn, P., & Shulman Rosengren, S. (1999). Attenuated familial adenomatous polyposis in a man with an interstitial deletion of chromosome arm 5q. *American Journal of Medical Genetics*, *86*(4), 321–324. https://doi.org/10.1002/(SICI)1096-8628(19991008)86:4<321::AID-AJMG4>3.0.CO;2-0

Pilati, C., Shinde, J., Alexandrov, L. B., Assié, G., André, T., Hélias-Rodzewicz, Z., … Laurent-Puig, P. (2017). Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *The Journal of Pathology*, *242*(1), 10–15. https://doi.org/10.1002/path.4880

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., … Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, *49*(10), 1476–1486. https://doi.org/10.1038/ng.3934

Popuri, V., Hsu, J., Khadka, P., Horvath, K., Liu, Y., Croteau, D. L., & Bohr, V. A. (2014). Human RECQL1 participates in telomere maintenance. *Nucleic Acids Research*, *42*(9), 5671–5688. https://doi.org/10.1093/nar/gku200

Quintana, I., Mejías-Luque, R., Terradas, M., Navarro, M., Piñol, V., Mur, P., … Valle, L. (2018). Evidence suggests that germline RNF43 mutations are a rare cause of serrated polyposis . *Gut*, *67*(12), 2230–2232. https://doi.org/10.1136/gutjnl-2017-315733

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature*, *505*(7483), 302–308. https://doi.org/10.1038/nature12981

Ramazzotti, D., Lal, A., Liu, K., Tibshirani, R., & Sidow, A. (2019). De novo mutational signature discovery in tumor genomes using SparseSignatures. *BioRxiv*, 384834.

https://doi.org/10.1101/384834

Ramos, P., Karnezis, A. N., Craig, D. W., Sekulic, A., Russell, M. L., Hendricks, W. P. D., … Trent, J. M. (2014). Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. *Nature Genetics*, *46*(5), 427–429. https://doi.org/10.1038/ng.2928

Ricciardone, M. D., Ozçelik, T., Cevher, B., Ozdağ, H., Tuncer, M., Gürgey, A., … Oztürk, M. (1999). Human MLH1 deficiency predisposes to hematological malignancy and neurofibromatosis type 1. *Cancer Research*, *59*(2), 290–293. Retrieved from http://cancerres.aacrjournals.org/content/59/2/290.abstract

Ried, T., Knutzen, R., Steinbeck, R., Blegen, H., Schröck, E., Heselmeyer, K., … Auer, G. (1996). Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, Chromosomes and Cancer*, *15*(4), 234–245. https://doi.org/10.1002/(SICI)1098-2264(199604)15:4<234::AID-GCC5>3.0.CO;2-2

Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., … Clevers, H. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, *556*(7702), 457–462. https://doi.org/10.1038/s41586-018-0024-3

Rohlin, A., Eiengård, F., Lundstam, U., Zagoras, T., Nilsson, S., Edsjö, A., … Nordling, M. (2016). GREM 1 and POLE variants in hereditary colorectal cancer syndromes. *Genes, Chromosomes and Cancer*, *55*(1), 95–106. https://doi.org/10.1002/gcc.22314

Rohlin, A., Zagoras, T., Nilsson, S., Lundstam, U., Wahlstr??m, J., Hult??n, L., … Nordling, M. (2014). A mutation in POLE predisposing to a multi-tumour phenotype. *International Journal of Oncology*, *45*(1), 77–81. https://doi.org/10.3892/ijo.2014.2410

Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E., & da Silva, I. T. (2017). signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, *33*(1), 8–16. https://doi.org/10.1093/bioinformatics/btw572

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, *17*(1), 31. https://doi.org/10.1186/s13059-016-0893-4

Rump, A., Benet-Pages, A., Schubert, S., Kuhlmann, J. D., Janavičius, R., Macháčková, E., … Klink, B. (2016). Identification and functional testing of ERCC2 mutations in a multi-national cohort of patients with familial breast- and ovarian cancer. *PLOS Genetics*, *12*(8), e1006248. https://doi.org/10.1371/journal.pgen.1006248

Schneppenheim, R., Frühwald, M. C., Gesk, S., Hasselblatt, M., Jeibmann, A., Kordes, U., … Siebert, R. (2010). Germline nonsense mutation and somatic inactivation of SMARCA4/BRG1 in a family with rhabdoid tumor predisposition syndrome. *American Journal of Human Genetics*, *86*(2), 279–284. https://doi.org/10.1016/j.ajhg.2010.01.013

Schubert, S. A., Ruano, D., Elsayed, F. A., Boot, A., Crobach, S., Sarasqueta, A. F., … van Wezel, T. (2017). Evidence for genetic association between chromosome 1q loci and predisposition to colorectal neoplasia. *British Journal of Cancer*, *117*(6), 1215–1223. https://doi.org/10.1038/bjc.2017.240

Schulz, E., Klampfl, P., Holzapfel, S., Janecke, A. R., Ulz, P., Renner, W., … Sill, H. (2014). Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. *Nature Communications*, *5*(May 2014), 5191. https://doi.org/10.1038/ncomms6191

Seguí, N., Mina, L. B., Lázaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., … Valle, L. (2015). Germline mutations in FAN1 cause hereditary colorectal cancer by impairing DNA repair. *Gastroenterology*, *149*(3), 563–566. https://doi.org/10.1053/j.gastro.2015.05.056

Seguí, N., Pineda, M., Navarro, M., Lázaro, C., Brunet, J., Infante, M., … Valle, L. (2014). GALNT12 is not a major contributor of familial colorectal cancer type X. *Human Mutation*, *35*(1), 50–52. https://doi.org/10.1002/humu.22454

Sharma, B., Handler, M., Eichstetter, I., Whitelock, J. M., Nugent, M. A., & Iozzo, R. V. (1998). Antisense targeting of perlecan blocks tumor growth and angiogenesis in vivo. *Journal of Clinical Investigation*, *102*(8), 1599–1608. https://doi.org/10.1172/JCI3793

Sharma, S., Stumpo, D. J., Balajee, A. S., Bock, C. B., Lansdorp, P. M., Brosh, R. M., & Blackshear, P. J. (2007). RECQL, a member of the RecQ family of DNA helicases, suppresses chromosomal instability. *Molecular and Cellular Biology*, *27*(5), 1784–1794. https://doi.org/10.1128/MCB.01620-06

Shinde, J., Bayard, Q., Imbeaud, S., Hirsch, T. Z., Liu, F., Renault, V., … Letouzé, E. (2018). Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*, *34*(19), 3380–3381. https://doi.org/10.1093/bioinformatics/bty388

Shiraishi, Y., Tremmel, G., Miyano, S., & Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLOS Genetics*, *11*(12), e1005657. https://doi.org/10.1371/journal.pgen.1005657

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., & Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, *67*(3), 177–193. https://doi.org/10.3322/caac.21395

Sill, H., Schulz, E., Steinke-Lange, V., & Boland, C. R. (2016). Correspondence: Reply to 'SEMA4A variation and risk of colorectal cancer'. *Nature Communications*, *7*(1), 10695. https://doi.org/10.1038/ncomms10695

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

Smith, C. G., Naven, M., Harris, R., Colley, J., West, H., Li, N., … Cheadle, J. P. (2013). Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. *Human Mutation*, *34*(7), 1026–1034. https://doi.org/10.1002/humu.22333

Smith, D. P., Rayter, S. I., Niederlander, C., Spicer, J., Jones, C. M., & Ashworth, A. (2001). LIP1, a cytoplasmic protein functionally linked to the Peutz-Jeghers syndrome kinase LKB1. *Human Molecular Genetics*, *10*(25), 2869–2877. https://doi.org/10.1093/hmg/10.25.2869

Snover, D. C., Ahnen, D. J., Burt, R. W., & Odze, R. D. (2010). Serrated polyps of the colon and rectum and serrated polyposis. In F. T. Bosman, F. Carneiro, R. H. Hruban, & N. D. Theise (Eds.), *WHO classification of tumours of the digestive system* (4th ed., pp. 160–165). Lyon: IARC.

Spier, I., Kerick, M., Drichel, D., Horpaopan, S., Altmüller, J., Laner, A., … Aretz, S. (2016). Exome sequencing identifies potential novel candidate genes in patients with unexplained colorectal adenomatous polyposis. *Familial Cancer*, 281–288. https://doi.org/10.1007/s10689-016-9870-z

Stenzinger, A., Allen, J. D., Maas, J., Stewart, M. D., Merino, D. M., Wempe, M. M., & Dietel, M. (2019). Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes, Chromosomes and Cancer*, *58*(8), 578–588. https://doi.org/10.1002/gcc.22733

Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. *Science*, *331*(6024), 1553–1558. https://doi.org/10.1126/science.1204040

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239),

719–724. https://doi.org/10.1038/nature07943

Sud, A., Kinnersley, B., & Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*, *17*(11), 692–704. https://doi.org/10.1038/nrc.2017.82

Takayama, T., Katsuki, S., Takahashi, Y., Ohi, M., Nojiri, S., Sakamaki, S., … Niitsu, Y. (1998). Aberrant crypt foci of the colon as precursors of adenoma and cancer. *New England Journal of Medicine*, *339*(18), 1277–1284. https://doi.org/10.1056/NEJM199810293391803

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., … Zhu, M. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, *35*(7), 899–907. https://doi.org/10.1002/humu.22537

Tan, V. Y. F., & Fevotte, C. (2013). Automatic relevance determination in nonnegative matrix factorization with the /spl beta/-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1592–1605. https://doi.org/10.1109/TPAMI.2012.240

Tanskanen, T., Gylfe, A. E., Katainen, R., Taipale, M., Renkonen-Sinisalo, L., Järvinen, H., … Aaltonen, L. A. (2015). Systematic search for rare variants in Finnish early-onset colorectal cancer patients. *Cancer Genetics*, *208*(1–2), 35–40. https://doi.org/10.1016/j.cancergen.2014.12.004

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., … Forbes, S. A. (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*, *3*, 92. https://doi.org/10.3389/fbioe.2015.00092

Taupin, D., Lam, W., Rangiah, D., McCallum, L., Whittle, B., Zhang, Y., … Cook, M. C. (2015). A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Human Genome Variation*, *2*(1), 15013. https://doi.org/10.1038/hgv.2015.13

Teer, J. K., & Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*, *19*(R2), R145–R151. https://doi.org/10.1093/hmg/ddq333

The Lancet. (2018). GLOBOCAN 2018: counting the toll of cancer. *The Lancet*, *392*(10152), 985. https://doi.org/10.1016/S0140-6736(18)32252-9

Thompson, E. R., Doyle, M. A., Ryland, G. L., Rowley, S. M., Choong, D. Y. H., Tothill, R. W., … Campbell, I. G. (2012). Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. *PLoS Genetics*, *8*(9), e1002894. https://doi.org/10.1371/journal.pgen.1002894

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), 110–124. https://doi.org/10.1038/nrg.2017.101

Toma, C., Díaz-Gay, M., Franch-Expósito, S., Arnau-Collell, C., Overs, B., Muñoz, J., … Castellví-Bel, S. (2019). Using linkage studies combined with whole-exome sequencing to identify novel candidate genes for familial colorectal cancer. *International Journal of Cancer*. https://doi.org/10.1002/ijc.32683

Toma, C., Shaw, A. D., Allcock, R. J. N., Heath, A., Pierce, K. D., Mitchell, P. B., … Fullerton, J. M. (2018). An examination of multiple classes of rare variants in extended families with bipolar disorder. *Translational Psychiatry*, *8*(1), 65. https://doi.org/10.1038/s41398-018-0113-y

Tomlinson, I. (2015). The Mendelian colorectal cancer syndromes. *Annals of Clinical Biochemistry*, *52*(6), 690–692. https://doi.org/10.1177/0004563215597944

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., & Issa, J.-P. J. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences*, *96*(15), 8681–8686. https://doi.org/10.1073/pnas.96.15.8681

Tuna, M., Knuutila, S., & Mills, G. B. (2009). Uniparental disomy in cancer. *Trends in Molecular Medicine*, *15*(3), 120–128. https://doi.org/10.1016/j.molmed.2009.01.005

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., Chapelle, A. d. l., Ruschoff, J., … Srivastava, S. (2004). Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *JNCI Journal of the National Cancer Institute*, *96*(4), 261–268. https://doi.org/10.1093/jnci/djh034

United Nations Development Programme. (2019). Human Development Index (HDI) | Human development reports. Retrieved from http://hdr.undp.org/en/content/human-development-index-hdi

Valle, L. (2017). Recent discoveries in the genetics of familial colorectal cancer and polyposis. *Clinical Gastroenterology and Hepatology*, *15*(6), 809–819. https://doi.org/10.1016/j.cgh.2016.09.148

Valle, L., de Voer, R. M., Goldberg, Y., Sjursen, W., Försti, A., Ruiz-Ponte, C., … Hemminki, K. (2019). Update on genetic predisposition to colorectal cancer and polyposis. *Molecular Aspects of Medicine*, *69*, 10–26. https://doi.org/10.1016/j.mam.2019.03.001

Valle, L., Vilar, E., Tavtigian, S. V, & Stoffel, E. M. (2019). Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. *The Journal of Pathology*, *247*(5), 574–588. https://doi.org/10.1002/path.5229

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Van Hoeck, A., Tjoonk, N. H., van Boxtel, R., & Cuppen, E. (2019). Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer*, *19*(1), 457. https://doi.org/10.1186/s12885-019-5677-2

Vasen, H. F., Mecklin, J. P., Khan, P. M., & Lynch, H. T. (1991). The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Diseases of the Colon and Rectum*, *34*(5), 424–425. https://doi.org/10.1007/bf02053699

Vasen, H. F., Watson, P., Mecklin, J. P., & Lynch, H. T. (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative Group on HNPCC. *Gastroenterology*, *116*(6), 1453–1456. https://doi.org/10.1016/s0016-5085(99)70510-x

Venkatachalam, R., Ligtenberg, M. J. L., Hoogerbrugge, N., Schackert, H. K., Görgens, H., Hahn, M.-M., … Kuiper, R. P. (2010). Germline epigenetic silencing of the tumor suppressor gene PTPRJ in early-onset familial colorectal cancer. *Gastroenterology*, *139*(6), 2221–2224. https://doi.org/10.1053/j.gastro.2010.08.063

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., … Bignami, M. (2017). A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine*. https://doi.org/10.1016/j.ebiom.2017.04.022

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., … Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, *319*(9), 525–532. https://doi.org/10.1056/NEJM198809013190901

Walsh, M. F., Ritter, D. I., Kesserwan, C., Sonkin, D., Chakravarty, D., Chao, E., … Plon, S. E. (2018). Integrating somatic variant data and biomarkers for germline variant classification

in cancer predisposition genes. *Human Mutation*, *39*(11), 1542–1552. https://doi.org/10.1002/humu.23640

Wang, Q., Lasset, C., Desseigne, F., Frappaz, D., Bergeron, C., Navarro, C., … Puisieux, A. (1999). Neurofibromatosis and early onset of cancers in hMLH1-deficient children. *Cancer Research*, *59*(2), 294–297. Retrieved from http://cancerres.aacrjournals.org/content/59/2/294

Watkins, J. A., Irshad, S., Grigoriadis, A., & Tutt, A. N. (2014). Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. *Breast Cancer Research*, *16*(3), 211. https://doi.org/10.1186/bcr3670

Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., … Laird, P. W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature Genetics*, *38*(7), 787–793. https://doi.org/10.1038/ng1834

Wellcome Trust Sanger Institute. (2019a). COSMIC: Catalogue of Somatic Mutations in Cancer - Mutational signatures (v2 - March 2015). Retrieved from https://cancer.sanger.ac.uk/cosmic/signatures_v2

Wellcome Trust Sanger Institute. (2019b). COSMIC: Catalogue of Somatic Mutations in Cancer - Mutational signatures (v3 - May 2019). Retrieved from https://cancer.sanger.ac.uk/cosmic/signatures

Weren, R. D. A., Ligtenberg, M. J. L., Geurts van Kessel, A., De Voer, R. M., Hoogerbrugge, N., & Kuiper, R. P. (2018). NTHL1 and MUTYH polyposis syndromes: two sides of the same coin? *The Journal of Pathology*, *244*(2), 135–142. https://doi.org/10.1002/path.5002

Weren, R. D. A., Ligtenberg, M. J. L., Kets, C. M., de Voer, R. M., Verwiel, E. T. P., Spruijt, L., … Hoogerbrugge, N. (2015). A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nature Genetics*, *47*(6), 668–671. https://doi.org/10.1038/ng.3287

Weren, R. D. A., Venkatachalam, R., Cazier, J. B., Farin, H. F., Kets, C. M., De Voer, R. M., … Kuiper, R. P. (2015). Germline deletions in the tumour suppressor gene FOCAD are associated with polyposis and colorectal cancer development. *Journal of Pathology*, *236*(2), 155–164. https://doi.org/10.1002/path.4520

Wimmer, K., Beilken, A., Nustede, R., Ripperger, T., Lamottke, B., Ure, B., … Kratz, C. P. (2017). A novel germline POLE mutation causes an early onset cancer prone syndrome mimicking constitutional mismatch repair deficiency. *Familial Cancer*, *16*(1), 67–71. https://doi.org/10.1007/s10689-016-9925-1

Win, A. K., Dowty, J. G., Cleary, S. P., Kim, H., Buchanan, D. D., Young, J. P., … Jenkins, M. A. (2014). Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology*, *146*(5), 1208-1211.e5. https://doi.org/10.1053/j.gastro.2014.01.022

Witkowski, L., Carrot-Zhang, J., Albrecht, S., Fahiminiya, S., Hamel, N., Tomiak, E., … Foulkes, W. D. (2014). Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nature Genetics*, *46*(5), 438–443. https://doi.org/10.1038/ng.2931

Wittschieben, J. P., Patil, V., Glushets, V., Robinson, L. J., Kusewitt, D. F., & Wood, R. D. (2010). Loss of DNA polymerase ζ enhances spontaneous tumorigenesis. *Cancer Research*, *70*(7), 2770–2778. https://doi.org/10.1158/0008-5472.CAN-09-4267

Wood-Trageser, M. A., Gurbuz, F., Yatsenko, S. A., Jeffries, E. P., Kotan, L. D., Surti, U., … Rajkovic, A. (2014). MCM9 mutations are associated with ovarian failure, short stature, and chromosomal instability. *American Journal of Human Genetics*, *95*(6), 754–762.

https://doi.org/10.1016/j.ajhg.2014.11.002

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., … Stratton, M. R. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, *378*(6559), 789–792. https://doi.org/10.1038/378789a0

Yan, H. H. N., Lai, J. C. W., Ho, S. L., Leung, W. K., Law, W. L., Lee, J. F. Y., … Leung, S. Y. (2017). RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. *Gut*, *66*(9), 1645–1656. https://doi.org/10.1136/gutjnl-2016-311849

Yang, L., Shi, T., Liu, F., Ren, C., Wang, Z., Li, Y., … Cheng, X. (2015). REV3L, a promising target in regulating the chemosensitivity of cervical cancer cells. *PLOS ONE*, *10*(3), e0120334. https://doi.org/10.1371/journal.pone.0120334

Yarchoan, M., Hopkins, A., & Jaffee, E. M. (2017). Tumor mutational burden and response rate to PD-1 inhibition. *New England Journal of Medicine*, *377*(25), 2500–2501. https://doi.org/10.1056/NEJMc1713444

Yehia, L., Niazi, F., Ni, Y., Ngeow, J., Sankunny, M., Liu, Z., … Eng, C. (2015). Germline heterozygous variants in SEC23B are associated with Cowden syndrome and enriched in apparently sporadic thyroid cancer. *American Journal of Human Genetics*, *97*(5), 661–676. https://doi.org/10.1016/j.ajhg.2015.10.001

Yu, C.-E., Oshima, J., Fu, Y.-H., Wijsman, E. M., Hisama, F., Alisch, R., … Schellenberg, G. D. (1996). Positional cloning of the Werner's syndrome gene. *Science*, *272*(5259), 258–262. https://doi.org/10.1126/science.272.5259.258

Yu, L., Yin, B., Qu, K., Li, J., Jin, Q., Liu, L., … Cao, K. (2018). Screening for susceptibility genes in hereditary non-polyposis colorectal cancer. *Oncology Letters*, *15*(6), 9413–9419. https://doi.org/10.3892/ol.2018.8504

Yurgelun, M. B., Kulke, M. H., Fuchs, C. S., Allen, B. A., Uno, H., Hornick, J. L., … Syngal, S. (2017). Cancer susceptibility gene mutations in individuals with colorectal cancer. *Journal of Clinical Oncology*, *35*(10), 1086–1095. https://doi.org/10.1200/JCO.2016.71.0012

Zhao, L., & Washington, M. (2017). Translesion synthesis: insights into the selection and switching of DNA polymerases. *Genes*, *8*(1), 24. https://doi.org/10.3390/genes8010024